160

The Behavior of PSYCHIATRIC PATIENTS

Quantitative Techniques for Evaluation

Edited by

EUGENE I. BURDOCK
Department of Psychiatry
New York University School of Medicine
New York, New York

ABRAHAM SUDILOVSKY
The Squibb Institute for Medical Research
Princeton, New Jersey
New York University School of Medicine
New York, New York

SAMUEL GERSHON
Lafayette Clinic
Department of Psychiatry
Wayne State University School of Medicine
Detroit, Michigan

MARCEL DEKKER, INC. New York and Basel

20 Some Contributions to Statistical Inference and Practice

W. Edwards Deming Consultant in Statistical Studies, Washington, D.C.

PURPOSE OF THIS CHAPTER

The aim of this chapter is to explain some points about statistical design and inference that have not yet found their way into textbooks. This is not a chapter on statistical techniques. It is more important, I believe, to try to understand the meaning of statistical results, than to try to write a condensation of techniques.

The importance of improved statistical design and inference can hardly be contemplated in this day of ever-increasing demands of reliability and safety in machines, apparatus, household appliances, including lawn mowers, swimming pools, and bicycles; and in safety of foods, drugs, and cosmetics, even in personal service such as the practice of medicine and law. In fact, suits for malpractice in medicine provide a good example of total failure on the part of physicians and lawyers to understand the distinction between (1) the inherent variation in performance of man and machine, and (2) variation from special causes, such as incompetence (1).

A regulatory agency, dealing in safety, reliability, or warranty, or with pollution, can hardly justify use of inefficient statistical tests and calculations. The statistical basis for branding some materials as carcinogenic, if published, would be interesting. One problem is that legislation on safety, reliability, and pollution is mostly devoid of operational meaning (cf. the section on operational definitions), without benefit of statistical thinking, and wide open to arbitrary interpretation.

TWO BASIC PROBLEMS

Statistical surveys, experiments, and trials may be classified into two basic types of problem: (1) enumerative and (2) analytic. This simple division in the planning of studies helps to avoid excessive costs and faulty inference (2).

ENUMERATIVE PURPOSE*

An enumerative study has for its aim an estimate of the number of units in a universe that belong to a specified class. The number of people 65 or over in some community, living at home, with severe dementia or with severe depression, by sex and age, and how they live, is an example. The number of days spent in the hospital and out of the hospital; all admissions; for the females in a region that were ever admitted to a mental hospital with a diagnosis of schizophrenia at first admission; all these are other examples. Further examples are: (1) Figures on the utilization of outpatient psychiatric services. (2) Prevalence of diabetes. (3) Assays of samples taken from a shipload of ore, or of wool, to estimate the dollar value of the shipload. The Bureau of Customs will collect duty on this estimate, if the material comes from abroad. (4) The Census for Congressional Representation in the United States is a prime example of an enumerative study. Congressional representation in an area depends on how many people are in it, not why they are there.

ANALYTIC PURPOSE

An analytic study has for its aim a basis for action on the cause system, to learn which treatment could improve the lot of people, or of patients of the future. Tests of varieties of wheat, insecticides, drugs, and manufacturing processes are examples of analytic studies. The results of the study will lead to choice of variety or treatment that will improve production or the lot of patients of the future. Tests for carcinogenic risk provide another example. Characteristics of the material tested are of interest only because they throw light on the cause system, through observation of the effects on the material tested under the environmental conditions of the study.

An experiment on rats to try to learn whether some specified food additive is harmful to rats is purely analytic. It in no way tells us how many rats there are, nor how many rats or people might, during the next decade, use the food additive or be affected by it.

*The reader might prefer to speak of enumerative studies as descriptive studies, and analytic studies as comparative studies—terms suggested by my friend, Professor S. Koller of the University of Mainz.

There is a simple criterion by which to distinguish between enumerative and analytic studies. A 100 percent sample of the frame provides the complete answer to the question posed for an enumerative problem, subject of course to the limitations of the method of investigation. In contrast, a 100 percent sample of a group of patients, section of land, or last week's product, industrial or agricultural, is still inconclusive in an analytic problem.

The purpose of a survey may be mainly enumerative, yet furnish information that is useful for important analytic purposes. The census is an example. It provides enumerative data by small areas concerning the inhabitants for use of demographers, sociologists, manufacturers, experts in marketing, doctors of public health, educators, and government. Comparison with previous censuses discloses trends in the age-sex-color composition of the population, migration, income, housing, and a host of other characteristics. School boards can prepare for more or fewer children in the primary schools of an area. Manufacturers of school desks can look ahead at the market for their product.

WHAT IS A FRAME?

An enumerative problem starts with a definition of the universe. The universe is all the material that we wish to describe.* This start, however, leaves us totally helpless until we find a way to take hold of portions of the universe. The frame is a means of access to the universe (3). A frame is composed of sampling units, any one of which may be selected for coverage. The totality of all the sampling units in the frame must cover enough of the universe to answer the essential questions about the universe. Without a frame, a complete coverage could neither be defined nor attempted, nor a sample less than 100 percent. In probability sampling, every sampling unit in the frame has an ascertainable probability of selection. Even a judgment selection of sampling units requires a frame. The frame may contain some blanks, often many blanks, not members of the universe.

In a survey of consumers of household soap, or of some household appliance, for example, the universe might be all the households in a region that provide a market for the article. The frame could be maps of small areas. Interviews in all the households in all these areas would cover the universe as defined.

For the census of population, the frame is a list of enumeration districts. Every square foot of the country lies in some area called an *enumeration district.* Every normal household lies in some enumeration district.

Materials in bulk present special problems, examples being a carload of wheat

*Following my friend, Frank Yates of Rothamsted, I use the word "material" to include the people, patients, business establishments, hospitals, institutions, accounts, land, insects, cows, crop of rice or wheat, and/or inventory that one wishes to describe or test.

or of ore, a pile of ore, a pile of coal. Most of these materials move sooner or later over a conveyor belt, which offers a chance to define workable sampling units.

Counts of fish in a lake and counts of wildlife also present fascinating problems in probability, not treated here. A related problem is estimation of the number of births and deaths in a region where the registration is incomplete (4).

WHEN IS A FRAME SATISFACTORY FOR AN ENUMERATIVE PURPOSE?

A frame is satisfactory (1) if it is composed of economical sampling units (a statistical question), and (2) if it covers enough of the universe to make the study of the frame worthwhile (a substantive question, for the sponsor of the study to decide).

In a study that a manufacturer of typewriters wished to carry out, a possible frame was a list of the 200,000 commercial concerns that employ 10 people or more. This frame did not include hospitals, schools, or libraries. Was the proposed frame satisfactory? The manufacturer decided in the affirmative: in his judgment, if he had information from all the concerns in the proposed frame, he would have information worth much more than its cost, even though it omitted 10 percent of his possible sales. He might later try to procure frames of schools, hospitals, and libraries, and study them, to encompass more of the universe of possible sales.

The manufacturer requires enumerative data in order to form a judgment about the magnitude of the market in important segments of business concerns that buy typewriters. This information will affect his product of the future and his methods of advertising and marketing. Future studies will have an analytic purpose, to test his methods.

THE EQUAL COMPLETE COVERAGE IN AN ENUMERATIVE STUDY

A result of the equal complete coverage (of a frame) is what one would get by studying, with the same method of investigation as specified for the survey, and with the same care, every sampling unit in the frame. The purpose of studying a sample less than a complete coverage is to derive estimates of what the results would be of the equal complete coverage.

There is an exception of importance. A complete coverage would find an extreme and unusual unit in the frame, if there be any. An example is a fraudulent account in a list of 100,000 accounts receivable. A sample of accounts less than 100 percent might miss this one account. Another is a pocket of diamonds buried in the ground. A sample of cores of earth to learn the amount of iron ore

in the field, and the dollar value of the field, might miss the diamonds. If any extreme units are thought to exist, they should be set off for separate treatment, preferably though not necessarily in advance of the sample coverage.

The equal complete coverage will contain nonresponse, illegible entries, missing entries, units missed, wrong interpretation of questions, wrong coding, and other problems, in the same proportion as found in the sample. The sample gives an estimate of the magnitude of these problems, as well as estimates of the number of people or of animals of various characteristics.

There are actual examples of the equal complete coverage. Every person enumerated in the census has a personal card or set of entries on a tape. The characteristics of his household will be on another card or tape. Most of the analytic work done by the census and by other government agencies and industrial companies is not made by use of the complete census but by samples of people and of households drawn from the complete census. The complete census is then an actual complete coverage of the samples drawn therefrom.

NONSAMPLING ERRORS

All data, whether obtained by a complete census or by a sample, are subject to various types of uncertainty. Nonsampling errors are always present, whether the study be enumerative or analytic. They belong to the equal complete coverage of the frame, not to be charged to use of samples. For example, nonresponse impairs a complete coverage as much as it impairs a sample less than 100 percent: likewise poor interviewing, poor supervision, method of investigation not suited to the purpose (e.g., the questionnaire, faulty techniques for examination of tumors), class intervals, and strata in the presentation of results that fail to display to the best advantage what is important in the results (5).

One may reduce uncertainties by recognizing their existence, measuring their magnitudes, and taking steps for improvement in future surveys. Sample design is an attempt to strike an economic balance between the different kinds of uncertainty. There is no point, for example, in reducing the sampling error of a result far below the level of other uncertainties.

THREE MAIN TYPES OF UNCERTAINTY

In my own work, I have found it helpful to divide the uncertainties of data into three types. Types I and II are nonsampling errors.

Type I

These are built-in deficiencies, or structural limitations, of the frame, questionnaire, or method of test.

Any reply to a question, or any record made by man or by an instrument, is only a response to a stimulus. What stimulus to apply is a matter of judgment. Deficiencies in the questionnaire or in the method of test may therefore arise from incomplete understanding of the problem or from unsuitable methods of investigation. Structural limitations are independent of the size or kind of sample. They are built in: a recanvass will not discover them, neither will calculation of standard errors nor other statistical calculations detect them. Some illustrations of uncertainty of type I are the following:

1. The frame for an enumerative study may fail to contain certain segments of the universe. The result will be an undercount of some of the total populations of the universe. This undercount may or may not be serious. (We saw an example with the typewriters.)
2. The questionnaire or method of test may fail to elicit certain information that is later found to be needed. The questionnaire may contain inept definitions, questions, and sequences. Detailed accounting will give results different from those given by mere inquiry about total expenditure of a family for some commodity; date of birth gives a different age from that given in answer to the simple question, How old are you? There may be differential effects of interviews depending on sex, race, and education of the interviewer.
3. Use of telephone or mail may yield results different from those obtained by personal interview.
4. One group of interviewers, hired, trained, and supervised to carry out a study, will produce results that differ from the results that another group will produce, hired, trained, and supervised by some other organization.
5. The date of the survey has an important effect on the answers in some studies.

Type II

Operational blemishes and blunders, for example, are:

1. Errors of a noncanceling nature (persistent omission of sampling units designated; persistent inclusion of sampling units not designated; persistent diagnosis of schizophrenia, no matter what be the ailment; drift of an instrument or of an observer).
2. Nonresponse leaves the result in doubt. There is no general limit on the percentage of response that is required. A small amount of nonresponse, like 5 percent, may seriously affect the results in some problems, whereas 20 percent nonresponse may not be serious in another problem.
3. Information supplied by coders for missing or illegible entries may favor high or low values.

4. There may be a single large error in the final result, such as a unique blunder, 86.8 for 68.8.

The magnitude of most of the uncertainties of type II can be detected and measured by use of statistical controls built into the procedure, by which random selections of sampling units are reinvestigated. Investment in statistical controls, in good practice, may well amount to 25 percent of the total budget for the study.

Type III

Uncertainty from random variation is type III. The sampling units in the frame are different from each other, hence repeated random samples drawn from the same frame will give different results by anybody's standards of observation. Besides, there are myriads of uncorrelated, nonpersistent, small accidental variations of a canceling nature that arise from inherent variability of investigators, supervisors, editors, coders, punchers, and other workers, and from random behavior of instruments. The t-test and other techniques are useful to construct a rule of behavior to detect faults in compliations in a frame, and errors in carrying out the specified procedure of selection of a sample therefrom, or in coding or tabulation.

STANDARD ERROR OF AN ESTIMATE

Type III is the only type of error for which man possesses a complete body of theory. The standard error of a result includes the combined effects of all kinds of random variation, including differences within and between investigators, supervisors, coders, etc. By proper design it is possible to get separate estimates of some of the components of variance.

SOME REMARKS ABOUT THE STANDARD ERROR OF AN ESTIMATE IN ENUMERATIVE STUDIES

1. The standard error is used the world over as a measure of precision, useful when the distribution of estimates is nearly normal. The smaller the standard error, the better the precision. Given an estimate of the standard error of a result, along with certain other statistical characteristics of components of the result, the statistician is able to calculate by theory the proportion of experiences in which a result calculated from repetitions of the sampling procedure would differ from the results obtained from a much larger sample, carried out by similar procedures with the same care and the same system and definitions.

2. A small standard error of a result in an enumerative study signifies (1) that the variation between repeated samples will be small, and (2) that the

result of the sample agrees well with the equal complete coverage. It usually tells little about uncertainties of type II and never anything about uncertainties of type I.

3. The expected payoff of an independent new sample would be zero. This is so because the result of any sample is as likely to be above as it is below the result of a complete study. The only thing that one can be sure of about a second sample is its cost. Thus, the expected payoff of a second sample is actually negative.

4. The expected payoff of a combination of the present sample and a new one of the same size, collected and processed in the same way, would favor the present sample because the combination would in large part be governed by the first sample. This is so because the correlation between the mean of a sample and a random half thereof is $\sqrt{.5} = .707$.

5. The expected payoff of a complete coverage, even if it could be carried out, now that this sample is in hand, would for the same reason be less than zero.

WHAT DO WE WISH TO LEARN FROM AN ANALYTIC PROBLEM?

What we need to know in a comparative study is whether the difference between two treatments, A and B, appears to be of material importance, economic or scientific, under the conditions of use in the future. This required difference we designate by D. Symbolically,

$$\text{Is } B \geqslant A + D?$$

That is, will B *in future trials* give a result bigger by the amount D than A will give? Treatment B could be modification of a production process, whereas treatment A is the standard way. Treatment B could be the effect of a food additive, whereas treatment A is no additive at all, except possibly distilled water.

The appropriate statistical design of an analytic study depends on the value of D, which in good experimental work is stated in advance. Its magnitude is the responsibility of the expert in the subject matter (doctor of medicine, pharmacologist, engineer, chemist, etc.), not of the statistician.

The statistical problem is one of estimation. Will process B turn out D more units per hour under the conditions to be met in service? Does compound X in hair spray cause 10 more cases of cancer in 10,000 women that use it than if they did not use it? (The statement that the compound could cause cancer is not an answer to the question because almost anything could cause cancer.) Does the experiment or series of experiments show that B is very likely better than A by the amount D? If so, then it would appear that treatment B ought perhaps to supplant treatment A. The doctor of medicine may decide, however, to await

confirmation in further trials. As he sees it, his patients are not the patients in the experiment: results might be different with his patients.

The farmer, learning from an agricultural experimental station that variety B yielded, in trials, more by the amount D bushels per acre than variety A yielded, may adopt variety B, or he may wait while other people try it. As he sees it, his farm has not the soil and climate of the experimental station. Moreover, all the environmental conditions could be different next year. It is next year that concerns him, not last year. On the other hand, he may get ahead of his friends by being the first one to try the new variety.

TWO KINDS OF ERROR IN CONCLUSIONS DRAWN FROM AN ANALYTIC STUDY

When we view an experiment or even a series of experiments from the standpoint of action, we perceive that there are two kinds of error to fall into. The action to take on the basis of an experiment, or on the basis of a series of experiments, will be:

1. Act too soon on a result, and regret later the decision made (wish that we held on to A).
2. Disregard the result, hold off adoption of what appears to be a better way, and regret later the decision made (wish that we had adopted B).

It is very simple to avoid consistently either one of these two errors. One may avoid 1 by never making a change, no matter what. One may avoid 2 by adopting new treatments and new methods as soon as results indicate the possibility of superiority of treatment B over treatment A.

Good science and good management consist in making, now and then, each error, but holding to a low level the net economic loss from both of them.

It is important to note that the statistician has only weak conditional formulas by which to minimize the net loss from both kinds of mistakes in an analytic problem. The reason is that one can not acquire enough empirical information by which to predict the environmental conditions of the future, or the performance of the treatments under these conditions.

DIFFERENCES IN ALLOCATION OF SAMPLE

Textbooks on statistical theory teach for estimation of a total population the importance of optimum allocation of the sample from the various strata in the frame. A familiar example is proportionate allocation by which (with two strata)

$$n_1 = nP_1$$
$$n_2 = nP_2 \qquad [1]$$

where n is the number of sampling units in the entire sample, n_1 and n_2 are the number of sampling units in strata 1 and 2, and P_1 and P_2 are the proportions of sampling units in the two strata. There is also Neyman allocation, sometimes useful, by which

$$n_1 = \frac{nP_1\sigma_1}{\bar{\sigma}_W}$$
$$n_2 = \frac{nP_2\sigma_2}{\bar{\sigma}_W} \qquad [2]$$

where

$$\bar{\sigma}_W = P_1\sigma_1 + P_2\sigma_2 \qquad [3]$$

σ_1 and σ_2 being the standard deviations between the sampling units in the two strata. In contrast, optimum allocation for an analytic study is

$$n_1 = n_2 \qquad [4]$$

regardless of how many people, patients, or acres of land belong in each stratum. Modification for different unit costs and variances in the strata, rarely encountered, requires only the factor $(\sigma_1/\sigma_2)(c_2/c_1)^{1/2}$ on the right of equation [4] (6).

We need not recognize strata at all in advance, or afterward, in an enumerative problem, although stratification in one form or another, either in advance or afterward, with use of ratio estimators or other estimation procedures with the aid of census information, might improve, with little added cost, the precision of some estimates.

In contrast, in an analytic problem, where the question is to discover whether and under what conditions two treatments A and B differ by the amount D or more, the expert in the subject matter may very wisely restrict the initial comparisons to strata where he would expect extremes of differences between treatments, such as areas of widely different climate and rainfall, or at the extremes of the severity of a disease. The experimenter bites off strata (areas, hospitals, patients) one at a time, until he establishes, in his judgment, the conditions under which B is superior to A by the amount D, or under which the difference is inconsequential. Caution is necessary, however.

> The mid-portion of pregnancy may be as vulnerable to environmental agents as early pregnancy . . . but the middle part is not included in drug-testing routines. (7)

In tests of food additives, conducted on, e.g., rats, homogeneous strains of rats are randomized for the various treatments. The more homogeneous the strain of rat, the better will be the chance to observe a difference between treatments, for those rats, but the less we learn about rats.

MORE ON THE USE OF JUDGMENT SAMPLES FOR ANALYTIC PURPOSES

In spite of the fact that we can at best arrange to carry out a comparison of treatments only on patients that are highly abnormal (usually patients that 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 experiments on two treatments appropriately randomized among the patients in a clinic indicate that the difference between two treatments is almost surely substantial (equal to D), then we have learned something: we may assert, with a calculable probability of being wrong, 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 use the treatments with patients like the ones tested, or raise wheat under the same environmental conditions. The establishment of a difference of economic or of scientific importance under any conditions chosen for convenience may constitute important new knowledge.

Randomization of treatments to patients within a clinic chosen for convenience, or to plots in an agricultural experiment station, eliminates the statistical problem of confounding treatments with patients, or treatments with plots, and opens the way for valid use of conditional statistical inference. The importance of randomization in the design of a judgment sample is thus obvious.

A WORD ON SIGNIFICANT DIFFERENCES

Any differences between A and B, however small, will be "significantly different" with enough cases. $P(\chi^2) \to 0$ as $n \to \infty$. Thus, no curve will fit data if we have enough data. Obviously, a criterion as soft as a sponge is not a useful method of inference.

There is nothing new here. A psychologist, a doctor of medicine, and two sociologists pointed out long ago the fact that a test of significance is not a useful method of inference (8-10).

In addition, tests of significance, t-test, chi-square, being symmetric functions, as statistical tests, all have the fault of losing the information in the order of the data (by time, or geographic location, or by seniority of investigator).

Efficient methods in analytic studies call for plots, run charts, scanning of m X n tables, $\bar{x}$- and R-charts. The theory of estimation will be useful to help to

decide whether $B - A \gtrless D$, provided the data show that the test method is stable or nearly so (11). The following quotation may be appropriate here (12): "When it is feasible, it is more informative to localize interaction in a few cells than it is to report a significant overall interaction with (R-1)(C-1) degrees of freedom."

It is fairly easy now to understand why it is that a general sample, spread all over the frame, as one would take it for an enumerative study, would be inefficient for an analytic study. Thus, to test two treatments in an agricultural experiment by randomizing the treatments in a sample of blocks drawn from a frame that consisted of all the arable blocks in the world would give a result that is almost useless, as a sample of any practical size would be so widely dispersed over so many conditions of soil, rainfall, and climate, that no useful inference could be drawn. The estimate of the difference B – A would be only an average over the whole world, and would not pinpoint the types of soil in which B might be distinctly better than A. An example comes from Koller (13).

> When the effect of strophantin (ouabain) on cardiac insufficiency is tested, it is not meaningful to estimate the average therapeutic effect for the total of cases of cardiac failure, for those patients already treated with digitalis respond badly to strophantin. It is more important to find out if there are contraindications than to estimate the structure of frequencies of the heterogeneous sub-groups and by this enumerate a general mean of the therapeutic criterion. (p. 237)
>
> For an etiological survey, tests in all areas of a population may be the wrong way to proceed. (p. 246, not verbatim)

ILLUSTRATIONS

I was invited to help in the interpretation of data on changes in chromosomes in identical twins of old age. Analysis of variance had showed that there was no significant difference between the twins in a pair. I asked for the original data, computed a certain characteristic from data, and plotted a simple chart (Fig. 1). It showed that measurements on the first member of a pair of twins gave slightly higher readings than measurements on the second member. The points are depicted in the accompanying chart. They show that there was something wrong in the method of measurement, even though the differences between members of pairs of twins was clinically unimportant. A simple statistical tool was effective where glamorous methods of cluster analysis with the computer were not. (Chart plotted from memory.)

Author example is offered in the following study of vision in boys and girls.

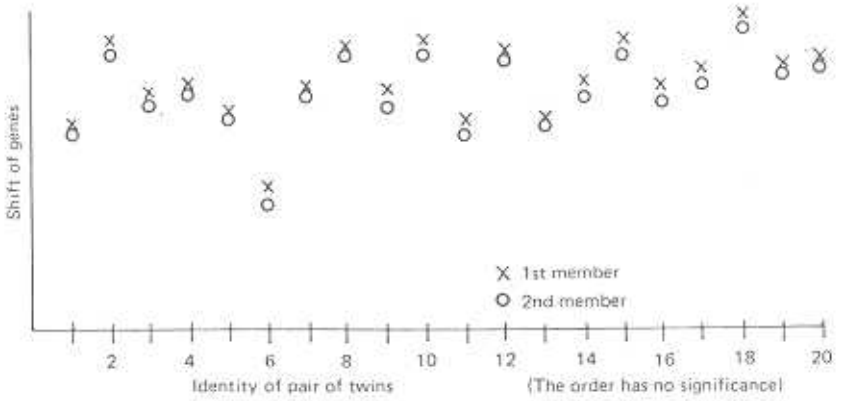

Figure 1 Chart depicting changes in chromosomes in identical twins of old age.

Boys 12-17 generally had better binocular distance acuity without correction than girls of that age in each of the four regions of the country. However, only in the Midwest and in the South were the differences . . . large enough to be statistically significant. (14)

U. S. 1966-1970

	20/20 or better				20/70 or poorer			
Sex	Northeast	Midwest	South	West	Northeast	Midwest	South	West
Boys								
Percent	72.1	70.0	80.0	74.5	16.7	18.2	10.6	13.8
Standard error	2.63	1.96	1.67	2.70	2.46	1.37	0.82	2.14
Girls								
Percent	66.3	60.9	72.6	66.7	18.3	23.6	13.3	21.2
Standard error	3.21	2.82	2.10	5.96	2.85	2.57	1.42	4.40

Source: From Ref. 14.

COMMENTS (15)

(1) The differences between boys and girls appear to be persistent from region to region, and to be substantial, of scientific importance, worthy of further study. (2) Examination of the detailed tables for the United States as a whole (not included here) give important evidence not mentioned in the article. Actually, more girls per 1,000 than boys per 1,000 at every age 12, 13, 14, 15, 16, and 17 have vision 20/17 and likewise 20/20, but more boys than girls have vision 20/15 and 20/12 or better. (3) What appears to be a higher percentage of boys in the accompanying table with vision 20/20 or better comes from the fortuitous consolidation and confounding of lopsided proportions at the different ages and levels of vision, of the kind just described. (4) These lopsided proportions may well be the most important result of the study, but the text bypasses this possibility. (5) The high proportions of both boys and girls in the South with vision 20/20 or better (not shown here), compared with the rest of the country, may have its origin in differences between the visions of black and white boys and girls, but the detailed tables do not show figures separately by color, possibly because of small samples for blacks. (6) Differences between examiners would, in my experience, be worth investigation, but the text gives no indication of how the boys and girls were allotted to the examiners, nor any summary of differences between examiners. (7) The standard errors shown in the table are meaningless; they apparently obscured the vision of the writer of the text.

LIMITATIONS OF STATISTICAL INFERENCE

All results are conditional on (1) the material from which came the units for test; (2) the method of investigation (the questionnaire or the test method and how it was used); (3) the people that carry out the interviews, measurements, or observations; in addition, (4) the results of an analytic study are conditional also on certain environmental states, such as the geographic locations of the comparison, the date and duration of the test, the soil, rainfall, climate, description and medical histories of the patients or subjects that took part in the test, the observers, the hospital or hospitals, duration of test, levels of radiation, range of voltage, speed, range of temperature, range of pressure, thickness (as of plating), number of flexures, number of jolts, maximum thrust, maximum gust, and maximum load.

The exact environmental conditions for any experiment will never be seen again. Two treatments that show little difference under one set of environmental circumstances or even within a range of conditions, may differ greatly under other conditions, other patients, other soils, other climate, etc. The converse may also be true: two treatments that show a large difference under one set of conditions may be nearly equal under other conditions.

There is no statistical method by which to extrapolate to longer usage of a drug beyond the period of test, nor to other patients, soils, climates, higher voltages, nor to other limits of severity outside the range studied. Side-effects may develop later on. Machinery that shows up well in a test that covers 3 weeks may cause grief and regret after a few months because of problems of maintenance. Economic conditions change, and upset predictions and plans. A competitor may step in with a new product, or put on a blast of advertising.

These are some of the reasons why information on an analytic problem can never be complete, and why computations by use of a loss function can only be conditional. The gap beyond statistical inference can be filled in only by knowledge of the subject matter (economics, medicine, etc.). It is easy to see the lack of caution in the following: "By the time these aircraft are in service, 9 of them will have completed 1250 hours of thorough testing under all conditions." This sentence was extracted from a letter to the author from one of the largest airlines in the country.

How could tests of the past cover all conditions to be met in the future? Upon receipt of this letter, I resolved immediately, for my own practice, to require the expert in the subject matter (engineer, lawyer) to specify in advance the ranges of stress under which the experiments would be conducted, as the results would be valid only within these ranges.

Presentation of results, to be optimally useful, and to be good science, must conform to Shewhart's rule: viz., preserve, for the uses intended, all the evidence in the original data (16).

The data of an experiment consist of much more than a mean and its standard deviation. The user of the results, in order to understand them, may require not only the original data, but also a description or reference to the method of investigation, the date, place, the duration of the test, a record of the faults discovered by the statistical controls, the amount of nonresponse by class, e.g., by income, and in some cases even the name of the observer. The reader of a paper is certainly entitled to know about any side-effects that were observed.

The statistician has an obligation, as architect of a study, to help his client to perceive in advance the limitations of any study that is contemplated, and to alter the design, if desirable, to meet the requirements.

An important question to ask before the plans for a study go too far is this: What will the results refer to? How do you propose to use them? The answer sometimes brings forth drastic modifications of the plans.

NEED FOR OPERATIONAL DEFINITIONS

Quantitative problems cannot be understood and cannot even be stated, nor can the effect of any treatment be evaluated without the aid of statistical theory and

methods. One cannot operationally define adjectives like improved, unimproved, reliable, safe, polluted, unemployed, on time (arrivals), equal (in size), round, random, tired, red, green, or any other adjective, for use in science, business, or in government, except in statistical terms. An operational definition is one that is communicable, one that people may use in conversation or in scientific dialogue or for a business contract. English, French, German, and Japanese cannot for scientific or business purposes convey the meaning of an adjective. A standard (as of safety, or of performance or capability), to have meaning for business or legal purposes, must be defined in statistical terms.

An illustration may be found in the definition of *improved* or *unimproved*, in reference to a patient. A consensus of two physicians usually turns out to be the opinion of one physician who dominates the other by superior position or greater experience. A better way is for each physician to note on paper, with respect to each patient, his opinion on whether the patient is improved or unimproved and to place each judgment in a sealed envelope for discussion later on. Comparison of difference, and discussion of a sample of agreements, will give a better picture of the patients, and will give the physicians a chance to learn from each other. It would also provide a comparison of (1) the variance between physicians within patients, with (2) the variance between patients. Thus, it would be possible to discover whether the two physicians are following the same criteria for improved and unimproved, and to learn need, when it exists, for sharper definitions.

It is commonly supposed that if two physicians understand their work, they would agree under very nearly all conditions (17). This is not so. Exact agreement is to be expected only at the extremes where there is under anybody's standards, no improvement, or definite improvement. Only under these two conditions can two physicians always agree. In the intermediate conditions, one physician will rightfully say that the patient is unimproved, while the other one, equally competent, using the same criteria, will say that the patient is improved. The probability of agreement on one patient is $p^2 + q^2$, which takes the value 1 at the extremes and reaches the minimum (½) where half the judgments would be 50:50 improved and unimproved. The symbol p is the proportion of a large number of qualified physicians that would declare the patient to be unimproved, $q = 1 - p$.

Most bickering between buyer and seller, and between departments of the same company, arise from failure of the buyer to state his specifications in meaningful terms. An illustration follows:

> *Houston.* John M. X. . . . said that he contracted with the Saudis to supply concrete housing forms, but he said the Saudis claimed after he arrived that the $400,000 deal included installing the forms and erecting the housing. (18)

REFERENCES

1. W. Edwards Deming. On some statistical aids toward economic production. *Interfaces 5*(4): (1975).
2. W. Edwards Deming. *Some Theory of Sampling.* Wiley, New York, 1950, Ch. 7; also, On probability as a basis for action. *Am. Stat. 29:*146-152 (1975).
3. F. F. Stephan. Practical problems of sampling procedure. *Am. Sociological Rev. 1:*569-580 (1936).
4. C. Chandrasekar and W. Edwards Deming. On a method of estimating birth and death rates and extent of registration. *J. Am. Stat. Assoc. 44:*101-115 (1949); D. S. Robson. Mark-recapture methods of population estimation. In *New Developments in Survey Sampling,* Norman L. Johnson and H. Smith (Eds.). Wiley-Interscience, New York, 1969, pp. 120-140.
5. W. Edwards Deming. *Sample Design in Business Research,* Wiley, New York, 1960.
6. W. Edwards Deming. *Sample Design in Business Research.* Wiley, New York, 1960, p. 358.
7. *New York Times,* 19 July, p. 24, 1975, quoting Dr. Andrew G. Hendricks.
8. Edwin G. Boring. Mathematical versus scientific significance. *Psychol. Bull. 15:*335-338 (1919).
9. Joseph Berkson. Tests of significance considered as evidence. *J. Am. Stat. Assoc. 37:*325-335 (1942).
10. Denton E. Morrison and Ramon E. Henkel. *The Significance Test Controversy,* Aldine Publishing Co., Chicago, 1970. Excellent for references and discussion.
11. John Mandel. The analysis of interlaboratory test data. *Standardization* 17-10 (1977); Joseph M. Cameron. *Measurement Assurance.* Document NBSIR, National Bureau of Standards, Washington, D.C., 1977, pp. 77-1240.
12. Cuthbert Daniel. Patterns in residuals in the two-way layout. Annual Princeton Conference in Applied Statistics, Princeton, N.J., December, 1976; *Technometrics 20*(4):385-395 (1978).
13. S. Koller. Use of non-representative surveys for etiological problems. In *New Developments in Survey Sampling,* Norman L. Johnson and H. Smith (Eds.). Wiley-Interscience, New York, 1969, pp. 235-246.
14. *Visual Acuity of Youths 12-17 Years.* National Center for Health Statistics, Series 11, No. 127, May, 1973.
15. These comments appeared in the *Am. Stat. 29:*146-152 (1975).
16. Walter A. Shewhart. *Statistical Method from the Viewpoint of Quality Control.* The Graduate School, Department of Agriculture, Washington, D.C., 1939, pp. 88, 89, 111, 135.
17. John Mandel, Mary N. Steele, and James Sharman. Analysis of interlaboratory studies of flammability of children's sleep-wear. *Standardization News 1*(5):9-12 (1973).
18. *Washington Post,* 17 May, p. A16, 1978.