145

On Probability As a Basis For Action

by

W. EDWARDS DEMING

Reprinted from
The American Statistician, Vol. 29, No. 4, 1975, pp. 146-152

On Probability As a Basis For Action*

W. EDWARDS DEMING**

Abstract

The aim of the author is improvement of statistical practice. The author distinguishes between enumerative studies and analytic studies. An enumerative study has for its aim an estimate of the number of units of a frame that belong to a specified class. An analytic study has for its aim a basis for action on the cause-system or the process, in order to improve product of the future. A fair price to pay for an inventory is an example of an enumerative study. Tests of varieties of wheat, insecticides, drugs, manufacturing processes, are examples of analytic studies: the choice of variety or treatment will affect the future out-turn of wheat, future patients, future product. Techniques and methods of inference that are applicable to enumerative studies lead to faulty design and faulty inference for analytic problems.

It is possible, in an enumerative problem, to reduce errors of sampling to any specified level. In contrast, in an analytic problem, it is impossible to compute the risk of making a wrong decision. The author provides a number of examples, and pleads for greater care in the writing and teaching of statistical theory and inference.

* * * * *

Aim and scope of this paper. The aim here is to try to contribute something to the improvement of statistical practice. The basic supposition here is that any statistical investigation is carried out for purposes of action. New knowledge modifies existing knowledge.

Urgent needs for statistical work. Challenges face statisticians today as never before. The whole world is talking about safety in mechanical and electrical devices (in automobiles, for example), safety in drugs, reliability, due care, pollution, poverty, nutrition, improvement of medical practice, improvement of agricultural practice, improvement in quality of product, break-down of service, break-down of equipment, tardy busses, trains, and mail, need for greater output in industry and in agriculture, enrichment of jobs. The consumer requires month by month ever greater and greater safety, and he expects better and better performance of manufactured articles. The manufacturer has the same problems in his purchases of materials, assemblies, machines, and use of manpower. He must, in addition, know more and more about his own product. What is due care in manufacturing? What is malpractice in medicine? Statistical work in consumer research is in a sorry state, more money being spent on it year by year, with ever worsening examples of practice and presentation.

These problems can not be understood and can not even be stated, nor can the effect of any alleged solution be evaluated, without the aid of statistical theory and methods. One can not even define operationally adjectives like reliable, safe, polluted, unemployed, on time (arrivals), equal (in size), round, random, tired, red, green, or any other adjective, for use in business or in government, except in statistical terms. A standard (as of safety, or of performance or capability) to have meaning for business or legal purposes, must be defined in statistical terms.

The label on a blanket reads "50 per cent wool." What does this mean? Half wool, on the average, over this blanket, or half wool over a month's production? What is half wool? Half by weight? If so, at what humidity? By what method of chemical analysis? How many analyses? The bottom half of the blanket is wool and the top half is something else. Is it 50 per cent wool? Does 50 per cent 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 select them? What criterion must the average satisfy? And how much variation between cuts is permissible? Obviously, the meaning of 50 per cent wool can only be stated in statistical terms. Mere words in English, French, or Japanese will not suffice. What means 80% butter fat in the butter that you buy?

Drastic changes in practice and in writing and in teaching are called for. As Shewhart said [18], the standards of knowledge and workmanship in industry and in public service are more severe than the requirements in pure science. He ought to have added that the requirements for statistical practice are also far more rigid than the requirements imposed on the teaching of statistics. It ought not to be that way, but it is. (More later on teaching.)

The frame, the universe, environmental conditions. A statistical study proceeds by investigation of the material in a frame [19]. The frame is an aggregate of identifiable tangible physical units of some kind, any or all of which may be selected and investigated. The frame may be lists of people, areas, establishments, materials, or of other identifiable units that would yield useful results if the whole content were investigated. It may be a lot of manufactured parts. Equally important in an analytic problem is a description of the environmental conditions that may affect the results (*vide infra*).

To facilitate exposition, we use a frame of N sampling units, numbered serially $1, 2, 3, \ldots, N$. However, there are circumstances in practice in which the size of

* I am indebted to many critics of earlier drafts of the manuscript for this paper; also to questions from the audience at lectures at a number of universities, including the Princeton meeting of the Biopharmaceutical Section of the American Statistical Association 4 Dec. 1974; the Universities of Mainz, Colorado, Wyoming, George Washington University, North Carolina, Inter-American Statistical Institute in Santiago de Chile.

** Consultant in Statistical Surveys, 4924 Butterworth Pl., Washington 20016.

the frame is indefinite, and a probability P of selection is applied to each sampling unit as it is presented [23]. A 100% sample (complete census) cannot be defined in the absence of a frame of N sampling units.

Enumerative studies and analytic studies contrasted. The distinction between enumerative and analytic studies is vital in the design of studies and in the interpretation of results.[1]

ENUMERATIVE: in which action will be taken on the material in the frame studied. The action to be taken on the frame depends purely on estimates or complete counts of one or more specific populations of the frame. The aim of a statistical study in an enumerative problem is descriptive. How many farms or how many people belong to this or that category? What is the expected out-turn of wheat for this region? How many units in the lot are defective? The aim is not to find out why there are so many or so few units in this or that category: merely how many.

Examples: 1. Data of Census-type: age, sex, education, occupation by area. 2. Figures on the utilization of out-patient psychiatric services. 3. Prevalence of diabetes. 4. Assays of samples taken from a shipload of ore, to estimate what the shipload is worth and to decide how much to offer for it. 5. Tests of manufactured product. (The Bureau of Customs will calculate also from the samples how much duty to pay, if the ore comes from abroad.) 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: in which action will be taken on the process or cause-system that produced the frame studied, the aim being to improve practice in the future. Examples: tests of varieties of wheat, comparison of machines, comparison of ways to advertise a product or service, comparison of drugs, action on an industrial process (change in speed, change in temperature, change in ingredients). Interest centres in future product, not in the material studied. Action: adopt Method B over A, or hold on to A, or continue the experiment.

There is a simple criterion by which to distinguish between enumerative and analytic studies. A 100 per cent 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 per cent sample of a group of patients, or of a section of land, or of last week's product, industrial or agricultural, is still inconclusive in an analytic problem. This point, though fundamental in statistical information for business, has escaped many writers.

The two types of problem call for different procedures of selection and calculation of estimates. For example, in an enumerative problem, where the aim is to estimate (e.g.) the number of females of age 20–29 in a given frame, we need not recognize strata at all in advance, nor afterward, though stratification in one form or another might improve the precision of the estimate without added cost.

In contrast, in an analytic problem, where the question is to discover where and under what conditions two treatments A and B differ by the amount D or more, we may very wisely restrict the initial comparisons to strata of widely different climate and rainfall, or at the extremes of the severity of a disease. Failure to perceive in advance by substantive knowledge which strata may react differently to A and B may greatly impair an analytic study. (Cf., the section, "Use of judgment-samples," infra.)

Two kinds of error in an enumerative problem. We can make either of two kinds of error in an enumerative problem. In the example mentioned above, we could:

1. Pay too much by the amount D or more for the ore tested.
2. Sell it for too little, by the amount D' or more.

First, before we can try to guard against one mistake or the other, we must decide on the error that we could tolerate. We should perhaps not mind paying \$500 too much, or if we were selling the ore would we mind receiving \$500 too little for it. We might accordingly be satisfied to set D and D' in this problem at \$500. The tolerance to aim at would depend on the economics involved [2], [11], [14], [15], [20].

Two kinds of error in an analytic problem. Here, we may:

1. Adopt Process B (replace A by B), and regret it later (wish that we held on to A).
2. Hold on to Process A, and regret it later (wish that we had adopted B).

The function of the statistician is to try to minimize the net loss from both kinds of mistakes, whether the problem be enumerative or analytic. He has formulas for doing so in an enumerative problem, but has only weak conditional formulas in an analytic problem.

Statisticians must face the fact that it is impossible to formulate a loss-function and minimize the net loss from the two errors that one can make in an analytic problem. We can accordingly not make consistently either error, even with the aid of statistical methods. The reason is that we cannot acquire enough empirical

[1] My friend and colleague Professor S. Koller of the University of Mainz suggests that enumerative studies might better be called descriptive studies, and that analytic studies that I deal with here might better be called comparative studies—that is, studies for comparing two treatments or two processes. I mention this in the interest of clarity, as the terms that he suggests may be helpful to many readers.

The distinction between enumerative and analytic studies has been in the air many years, but has successfully eluded most books and most teaching. It was clearly recognized, without use of symbolic terms, by Harold F. Dodge and Harry Romig in their *Sampling Inspection Tables* (Wiley, 1944), and by the Statistical Research Group headed by W. Allen Wallis at Columbia University; see *Sampling Inspection* (McGraw-Hill, 1947), pp. 183–184. See also reference [6].

data to predict the environmental conditions of the future, nor the performance therein.

Comparison of two treatments is not a uniformity test by which to estimate the variance between plots within blocks subjected to the same treatment, or the variance between patients under the same treatment. It is not an investigation to ask whether the data conform to some specific genetic law of inheritance, which might generate the ratio 3:1 for light hair and dark hair in the offspring.

Limitations of statistical inference. All results are conditional on (a) the frame whence came the units for test; (b) the method of investigation (the questionnaire or the test-method and how it was used); (c) the people that carry out the interviews or measurements. In addition (d), 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, 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 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. Problems of maintenance of machinery that show up well in a test that covers three weeks may cause grief and regret after a few months. A competitor may step in with a new product, or put on a blast of advertising. Economic conditions change, and upset predictions and plans. 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, chemistry, engineering, psychology, agricultural science, etc.), which may take the formality of a model [12], [14], [15]. These admonitions seem to be ignored in books for decisions in business, the very place where they are most needed [16]. It is easy to see the fallacy in the following paragraph:

> By the time these aircraft are in service, 9 of them will have completed 1250 hours of thorough testing under all conditions.—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 will be conducted, and to explain that the results will be valid only within these ranges.

In work with a railway that hauls pellets of iron ore, tests over two years indicated small correlation between the number of loaded cars in a train and the average net weight of pellets per car in the train: also little effect of the weather, month to month, on the loading. This information turned out to be useful in the judgment of the engineers and accountants because they were confident that about the same correlation would hold the next year and the year after. Statistical theory could not predict the correlation.

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 [18].

The data of an experiment consist of much more than a mean and its standard deviation. In fact, not even the original observations constitute all the data. The user of the results, in order to understand them, may require 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, and in some cases, even the name of the observer [17]. An example of presentation that covers some of these points appears in an article by Butterworth and Watts [5].

> No side effects were observed among patients in any of the three study-groups, and no abnormal laboratory values were recorded over the 3-week trial.

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.

What do we need? 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,

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

That is, will B be better than A by the amount D IN FUTURE TRIALS? Will Process B turn out D more units per hour under the conditions to be met in the factory?

The appropriate statistical design depends on the value of D, which must be stated in advance. Its magnitude is the responsibility of the expert in the subject-matter.

The problem is one of estimation. What is the magnitude of $B - A$?

It is important to remember that the mean, the variance, the standard error, likelihood, and many other functions of a set of numbers, are symmetric. Interchange of any two observations x_i and x_j leaves unchanged the mean, the variance, and even the distribution itself. Obviously, then, use of variance and elaborate methods of estimation buries the information contained in the order of appearance in the original data, and must therefore be presumed inefficient until cleared.

Pencil and paper for construction of distributions, scatter diagrams, and run charts to compare small groups and to detect trends, are more efficient methods of estimation than statistical inference that depends on variances and standard errors, as the simple techniques preserve the information in the original data. In fortunate circumstances (normal estimates, independence, absence of patterns), and when the whole study went off as intended, one may indeed summarize the results of comparisons as confidence intervals or fiducial intervals, making use of standard errors. But these circumstances require demonstration by simple methods of pencil and paper [1], [7], [21].

We admit with Sir Winston Churchill that it sometimes pays to admit the obvious: we do not perform an experiment to find out if two varieties of wheat or two drugs are equal. We know in advance, without spending a dollar on an experiment, that they are not equal.

The difference between two treatments or between two areas or two groups of people, will show up as "significantly different" if the experiment be conducted through a sufficient number of trials, even though the difference be so small that it is of no scientific or economic consequence.

Likewise, tests of whether the data of a survey or an experiment fit some particular curve is of no scientific or economic consequence. $P(\chi^2) \to 0$ for any curve as the number of observations increases. With enough data, no curve will fit the results of an experiment. The question that one faces in using any curve or any relationship is this: How robust are the conclusions? Would some other curve make safer predictions?

Statistical significance of B over A thus conveys no knowledge, no basis for action [3], [4], [6], [8], [9], [13], [22].

Use of judgment-samples. Statisticians must face some facts about judgment-samples.

1. Use of a judgment-sample of material and environmental conditions for an enumerative study (e.g., by which to estimate the frequency of error of a certain type in a class of accounts) is worth no more than the reputation of the man that signs the report. The reason is that there is no way except by his judgment to set limits on the margin of uncertainty of the estimate. Probability samples have an advantage in an enumerative problem, as they remove one important area of doubt; they enable one to evaluate the uncertainty in a result that arises from (a) the myriads of independent chance causes of variation; (b) the variance between investigators; (c) interactions between investigators and sampling units; (d) the effects of the possible main flaws in execution.

2. Use of judgment-samples is hardly ever necessary in an enumerative problem. It may seem that exceptions occur in a pile of coal or in a shipload of ore where one can only take samples from exposed portions. There are usually ways around such difficulties, namely, to draw samples off the conveyor-belt while the coal or ore is being loaded or unloaded.

3. In contrast, much of man's knowledge in science has been learned through use of judgment-samples in analytic studies. Rothamsted and other experimental stations are places of convenience. So is a hospital, or a clinic, and the groups of patients therein that we may examine.

4. 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 amongst the patients in a clinic indicate that the difference 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 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 contribution is incomplete, but it is nevertheless a contribution.

5. The last paragraph brings up the importance of randomization and theories of experimental design in the use of judgment-samples. Randomization within the blocks in an area chosen for convenience (for trials of wheat), or of patients (for comparison of treatments), removes an important area of doubt. Under fortunate conditions, randomization of treatments within a selected stratum justifies the use of probability for conditional inferences. Theories of experimental design help to minimize the variances for a given allowable cost. But every inference (conclusion) based on the results is conditional, no matter how efficient be the design of experiment.

6. We have already observed that selection of widely different strata may be the most efficient approach for a comparison of treatments. One may bite off strata (areas, hospitals, patients) one at a time, as results seem to indicate, until he has, in his judgment, covered

enough strata and conditions to establish the areas and conditions under which the superiority of B over A is equal to or greater than D, or in which the difference is inconsequential. Omission of a stratum of special interest may impair an experiment. Example:

> 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.—New York Times, 19 July 1975: page 24, quoting Dr. Andrew G. Hendrickx.

It is fairly easy now to understand why it is that a probability sample of a whole frame 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 nigh 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 pin-point the types of soil in which B might be distinctly better than A. An exception would occur if treatment B turned out everywhere to be substantially superior to A. I am only reinforcing Koller: [10]

> 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.
>
> For an etiological survey representativeness of the total population is not an important criterion to distinguish between good and bad studies.

Criticisms of teaching. We are ready now to offer some specific criticisms on the teaching of statistics, in the hope that they may help to improve statistical practice in the future.

Example 1, composite example, extracted from a number of textbooks in mathematical statistics.

> Nineteen chutes were tested for time of burning, 10 in Batch A, 9 in Batch B. . . . $t = 1.40$. As $t_{.025} = 1.96$, it follows that the null hypothesis (that the two populations are identical) can not be rejected against the alternative $\mu_1 \neq \mu_2$ at the level of significance $\alpha = 0.05$.

Comments. What do the results refer to? What is a population? What two populations are not identical? Does the author of the book refer to the two batches of chutes already produced and tested, or does he mean the processes by which chutes will be produced? Only the manufacturing processes could be of concern, as the chutes tested (and burned up) will never be used again. What about a reference to the method of test? None is cited. The problem should be stated as an analytic problem in which the question is whether the economics of production, marketing, and consumer preference would justify a choice between processes A and B.

If a statistician in practice were to make a statement like the one quoted, he would lose his job summarily, or ought to.

To extract the information from so costly an experiment, I should wish to have in hand the original data, to be able to plot the time of burning of every chute tested, in the order of test. Were A and B alternated? If not, why not? Students need encouragement to think and to ask questions of the data, but how could they here? Missing from the text are the original figures and a description of how the tests were conducted (what Shewhart called the "data of the experiment" [17]).

Are students made aware that standard errors and statistical tests ignore all these questions? that standard errors bury a lot of essential information?

Under the usual teaching, the trusting student, to pass the course, must forsake all the scientific sense that he has accumulated so far, and learn the book, mistakes and all.

Example 2 from my friend Dr. Edward C. Bryant. While he was at the University of Wyoming, someone came in from the Department of Animal Husbandry to announce to him an astounding scientific discovery—the fibres on the left side of sheep, and those on the right side, are of different diameter. Dr. Bryant asked him how many fibres he had in the sample: answer, 50,000. This was a number big enough to establish significance. But what of it? Anyone would know in advance, without spending a dollar, that there is a difference between fibres of the left side and the right side of any sheep, or of n sheep combined. The question is whether the difference is of scientific importance.

Example 3. The Panel on Statistics distributed at the meeting of the American Statistical Association in Montreal in August 1972, in the pamphlet INTRODUCTORY STATISTICS WITHOUT CALCULUS, the following statement (page 20).

> A basic difficulty for most students is the proper formulation of the alternatives H_0 and H_1 for any given problem and the consequent determination of the proper critical region (upper tail, lower tail, two-sided). (Here H_0 is the hypothesis that $\mu_1 = \mu_0$; H_1 the hypothesis that $\mu_1 \neq \mu_0$.)

Comment. Small wonder that students have trouble. They may be trying to think.

Example 4, taken from "Visual Acuity of Youths 12–17 Years," National Center for Health Statistics, Series 11, No. 127, May 1973.

> 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.

U. S. 1966–70

Sex	20/20 or better				20/70 or poorer			
	North-east	Mid-west	South	West	North-east	Mid-west	South	West
Boys								
Per cent	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								
Per cent	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

Comments. (a) 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. (b) Examination of the detailed tables for the U. S. as a whole (not included here) give evidence in conflict with the conclusion quoted. Actually, more girls per 1000 than boys per 1000 at every age 12, 13, 14, 15, 16, 17, have vision 20/17 and likewise 20/20, but more boys than girls have vision 20/15 and 20/12 or better. (c) 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 lop-sided proportions at the different ages and levels of vision, of the kind just described. (d) These lop-sided proportions may well be the most important result of the study, but the text by-passes this possibility. (e) 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. (f) 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. (g) The standard errors shown in the table are meaningless; they apparently obscured the vision of the writer of the text.

More on the teaching of statistics. Little advancement in the teaching of statistics is possible, and little hope for statistical methods to be useful in the frightful problems that face man today, until the literature and classroom be rid of terms so deadening to scientific enquiry as null hypothesis, population (in place of frame), true value, level of significance for comparison of treatments, representative sample. There is no true value of any concept that is measured. There may be, of course, an accepted operational definition (questionnaire, method of measurement) and an accepted value—accepted until it is replaced with one that is more acceptable to the experts in the subject-matter [6], [12], [18].

Here are three suggestions to replace topics that should be thrown out. First, non-sampling errors: their detection and measurement by statistical controls, and their possible effects on uses of the results [6]. Second, the contrast between enumerative and analytic studies, their purposes and contrast in design and analysis. This would automatically bring in the use of judgment-samples where they are indicated for best efficiency. Third, every student should try his hand at statistical inference, given a set of original data, together with the necessary non-statistical information about the environmental conditions of the experiment or survey. Cross-examination by other members of the class would teach a student to be careful.

Students of statistics need some teachers that are engaged in practice. What would happen in medicine if medical students studied surgery and internal medicine from physicians none of whom had ever been in practice?

A teacher that gets involved in statistical problems has a basis for making a choice of what theory to teach, and he has illustrations of his own for the classroom and for his book, and he will understand the illustrations. His teaching inspires students to think.

We also need some teachers of theory that are not in practice. The student can only learn theory. To learn theory, though, most students require examples of good practice and examples of bad practice, with explanation, in terms of theory, of what was right and what was wrong about the procedure. Faultless, skillful teaching of statistics is unhappily too often undone by examples of design and inference that mislead the student, as I have tried to illustrate here.

Unfortunately, involvement in a problem carries responsibilities. The statistician in practice must write a report for management or for legal purposes, on the statistical reliability of the results, what they may mean and what they don't mean. Such a report will state the possible margins of uncertainty from sampling variation and from operational blemishes big and little discovered in the controls, the nonresponse, illegible entries, missing entries, inconsistencies found in the coding, with a careful statement of the conditions of the experiment (duration, locality, reasons for choice thereof, voltage, range of stress, etc.), the method of measurement or the questionnaire, and the difficulties encountered. (An example is in reference [6].) Involvement in a problem means the possibility of facing a board of directors, or facing cross-examination. It means tedious work, such as studying the data in various forms, making tables and charts and re-making them, trying to use and preserve the evidence in the results and to be clear enough to the reader: to endure

disappointment and discouragement. Desultory advice on possible ways to attack a double integral does not constitute involvement in a problem.

REFERENCES

[1] Anscombe, F. J.: "Graphs in Statistical Analysis," *The American Statistician*, Feb. 1973, vol. 27, No. 1.

[2] Blythe, Richard H.: "The economics of sample-size applied to the scaling of sawlogs," *The Biometrics Bulletin*, Washington, vol. 1, 1945: pp. 67–70.

[3] Berkson, Joseph: "Tests of significance considered as evidence," *Journal of the American Statistical Association*, vol. 37, 1942: pp. 325–335.

[4] Boring, Edwin G.: "Mathematical versus scientific significance," *Psychological Bulletin*, vol. 15, 1919: pp. 335–338.

[5] Butterworth, Alfred T., and Watts, Robert D.: "Double blind comparisons," *Psychomatics* xv, 1974: pp. 85–87.

[6] Deming, W. Edwards: *Some Theory of Sampling* (Wiley, 1953; Dover 1960), Ch. 5. An example of a statistical report appears in this book on pages 159–163. See also "Boundaries of statistical inference," being Chapter 31 in the book by Norman L. Johnson and Harry Smith: *New Developments in Survey Sampling* (Wiley, 1969).

[7] Ehrenberg, A. S. C.: *Data Reduction* (Wiley, 1975), pp. vii, 88, 350 in particular.

[8] Erhardt, Carl: "Statistics, a trap for the unwary," *Obstetrics and Gynecology*, vol. 14, Oct. 1959: pp. 549–554.

[9] Hansen, Morris H., and Deming, W. Edwards: "On an important limitation to the use of data from samples," *Bulletin de l'institute international de statistique*, Bern 1950: pp. 214–219.

[10] Koller, S.: "Use of non-representative surveys for etiological problems," a chapter in the book edited by Norman L. Johnson and Harry Smith: *New Developments in Survey Sampling* (Wiley-Interscience, 1969): pp. 235–246.

[11] Kruskal, William H.: "Tests of significance," *International Encyclopedia of the Social Sciences*, vol. 14, pp. 238–250. "Chromosomal effect and LSD: samples of 4," *Science*, vol. 162, 27 Dec. 1968, pp. 1508–1509.

[12] Lewis, C. I.: Mind and the World-Order (Scribner, 1929), p. 238.

[13] Morrison, Denton E., and Henkel, Ramon E.: *The Significance Test Controversy* (Aldine Publishing Company, 1970). Excellent for references and discussion.

[14] Savage, I. Richard: *Statistics: Uncertainty and Behavior* (Houghton Mifflin, 1968), Ch. 4.

[15] Schlaifer, Robert: *Probability and Statistics for Business Decisions* (McGraw-Hill, 1959), Ch. 33.

[16] Schlaifer, Robert: loc. cit., p. 491.

[17] Shewhart, Walter A.: *Statistical Method from the Viewpoint of Quality Control* (Graduate School, Department of Agriculture, 1939), p. 120.

[18] Shewhart, Walter A.: loc. cit., pp. 88, 89, 111, 135.

[19] Stephan, F. F.: American Sociological Review, vol. 1, 1936: pp. 569–580.

[20] Törnqvist, Leo: "An attempt to analyze the problem of an economical production of statistical data," Nordisk Tidsskrift for Teknish Økonomi, vol. 37, 1948: pp. 263–274.

[21] Tukey, John: "Some graphic and semigraphic displays," being Ch. 18 in *Statistical Papers in Honor of George W. Snedecor*, edited by T. A. Bancroft (Iowa State University Press, 1972). *Exploratory Data Analysis* (Addison-Wesley Press, 1975).

[22] Wolfowitz, J.: "Remarks on the theory of testing hypotheses," *The New York Statistician*, March 1967, vol. 18, No. 7.

[23] Yates, Frank: *Sampling Methods and Surveys* (Griffin, 1949, 1960), Ch. 3. Jolly, G. M., "Explicit estimates from capture-recapture data," *Biometrika*, vol. 52, 1965: pp. 225–247. W. Edwards Deming: "On variances and estimators of a total population under several procedures of sampling," *Methods in Biometry*, Festschrift für Dr. A. Linder (Birkhäuser, Basel, 1975).

147

REPORT TO MANAGEMENT

By

W. EDWARDS DEMING

CONSULTANT IN STATISTICAL STUDIES

WASHINGTON 20016
4924 BUTTERWORTH PLACE

TEL. (202) EMERSON 3-8552

REPRINTED FROM
QUALITY PROGRESS
VOL. 5, NO. 7
JULY 1972, PAGE 2

REPORT TO MANAGEMENT

In response to the January 1972 Viewpoints column on process capability, ASQC Honorary Member Dr. W. Edwards Deming sent us the following "report" composed of extracts from a report to the management of a large company. In addition to its being on process capability, it covers so many other items of interest to quality controllers in general that we present it here in its entirety.

W. E. DEMING
Categorizing troubles.

This report is written at your request after study of some problems that you are having with production, high costs and variable quality, which altogether, as I understand you, have been the cause of considerable worry to you about your competitive position. Please note that I write as a statistician who sees the statistical method as a system of service to science and to industry. I am not a consultant in management. As a statistician in practice, however, I work with management on many types of problems, including statistical logic in the management of quality. Thus I learn what some management problems are and how statistical methods can help.

By quality control, I mean use of statistical methods to aid design and test of product, specifications and tests of materials, aids to production workers, measurement of the effects of common (environmental) causes, meaningful job descriptions and specifications based on the capability of the process, consumer research, sales, inventory, inventory-policy, maintenance of equipment and many other problems of management.

My opening point is that no permanent impact has ever been accomplished in quality control without understanding and continued nurture of top management. No short-cut has been discovered. In my opinion, failure of your own management to accept and act on their responsibilities in quality control is one cause of your trouble, as further paragraphs will indicate in more detail.

What you have in your company, as I see it, is not quality control, but guerrilla sniping — no organized system, no provision nor appreciation for the statistical control of quality as a system. You have been running along with a fire department that hopes to arrive in time to keep fires from spreading. Your quality control department has done its duty, as I understand, if they discover that a carload of finished product might cause trouble (even legal action) if it went out. This is important, but my advice is to build a system of quality control that will reduce the number of fires in the first place. You spend money on quality control, but ineffectively.

You have a slogan, posted everywhere. I wonder how anyone could live up to it. By every man doing his job better? How can he, when he has no way to know what his job is nor how to do it better? Exhortations and platitudes are not effective instruments of improvement in today's fierce competition, where a company must compete across national boundaries. Something more is required.

A usual stumbling block most places (except in Japan, I believe, where they had the benefit of a better start, and a willingness of top management to learn and stay interested) is management's supposition that quality control is something that you install, like a new dean or a new carpet.

Another roadblock is management's supposition that the production workers are responsible for all trouble: that there would be no problems in production if only production workers would do their jobs in the way that they know to be right. Man's natural reaction to trouble of any kind in the production line is to blame the operators. Instead, in my experience, most problems in production have their origin in common (environmental) causes, which only management can reduce or remove. For best economy, the production worker is held responsible to maintain statistical control of his own work. To ask him to turn out no defectives may be costly and the wrong approach. The QC Circle movement in Japan gives to production workers the chance to

move on certain types of common causes, but the QC Circle movement is in Japan, not here.

Causes of trouble may be subsumed under two categories: common (environmental) and special (local). Common causes are called common because they affect equally all workers in a section. They are faults of the system. They stay there until removed by management. Their combined effect can be evaluated. Individual common causes can usually be isolated by experiment. Special cause can be corrected on statistical signal by the production worker himself. They are special because they are specific to a local condition. The operator's judgment by itself without statistical signals is hazardous.

Confusion between common causes and special causes — a failure of management — is one of the most costly mistakes of industry administration, and public administration as well. Confusion between these two causes leads to frustration at all levels and to actual increase in variability and cost of product — exactly contrary to what is needed.

Fortunately, confusion between the two sources of trouble (common or environmental causes, and special causes) can be eliminated with almost unerring accuracy. Simple statistical methods distinguish between the two types of cause, and thus point the finger at the source and at the level of responsibility for action. Simple statistical charts tell the operator when to take action to improve the uniformity of his work, and when to leave it alone. Moreover, the same simple statistical tools can be used to tell management how much of the proportion of defective material is chargeable to common (environmental) causes, correctible only by management.

Thus, with simple data, it is possible and usually not difficult to measure the combined effect of common causes on any operation. This I pointed out in my paper "On Some Statistical Logic in The Management of Quality," which I delivered at the All India Congress on Quality Control held in New Delhi, 17 March 1971.

"We rely on our experience," is the answer that came from the quality manager in a large company recently, when I enquired how they distinguish between the two kinds of trouble (special and environmental) and on what principles. Your own people gave me the same answer, at first.

This answer is self-incriminating — a guarantee that the company will continue to have about the same amount of trouble. There is a better way now. Experience can be cataloged and put to use rationally only by application of statistical theory. One function of statistical methods is to design experiments and to make use of relevant experience in a way that is effective. Any claim to use of experience without a plan based on theory is disquise for rationalization of a decision that has already been made.

In connection with special causes, I find in your company no provision to feed back to the production worker information in a form that would indicate (a) when action on his part would be effective in helping to meet his specifications, and (b) when he should leave his process as it is. Special causes can be detected only with the aid of proper statistical techniques.

The production worker himself may in most cases plot the statistical charts that will tell him whether and when to take action on his work. He must, of course, be taught.

Be it noted, though, that statistical techniques for special causes alone will be ineffective and will fizzle out unless management has taken steps to remove the common (environmental) causes of trouble that make it impossible for the production worker to turn out good work. Failure of management to take this initial step, before teaching the production worker how to detect his own special causes, accounts for failure of the so-called control chart method; it simply will not solve all the problems of quality.

The benefit of this communication with the worker, by which he perceives a genuine attempt on the part of management to show him what his job is, and to hold him responsible for what he himself can govern, and not for the sins of management, is hard to overestimate.

Moreover, there is a further elevation of morale when the worker perceives that management is doing something about common causes, and accepting some of the blame for trouble.

Statistical aids to the production worker will require your company to acquire some statistical knowledge and do a lot of planning.

What is the production worker's job? Is it to turn out no defectives (which makes him responsible, not just for his own work, but for the machinery and for the material that comes to him from previous operations, or is it to run his operation economically? The two aims are too often incompatible. Statistical methods show up this dilemma and provide feasible solution.

There is no excuse today to hand to a worker specifications that he cannot meet, nor to put him in a position where he cannot tell whether he has met them. Your company fails miserably here.

When a process has been brought into a state of statistical control (special causes weeded out), it has a definite capability, expressible as the economic level of quality for that process.

The only specifications with meaning are those fixed by the capability of the process. The specifications that a process in control can meet are obvious. There is no process, no capability, and no meaningful specifications, except in statistical control.

Tighter specifications can be realized only by reduction or removal of some of the common causes of trouble, which means action on the part of management. A production worker, when he has reached statistical control, has put into the process all that he has to offer. It is up to management to provide better uniformity in incoming materials, better uniformity in previous operations, better setting of the machine, better maintenance, change in the process, change in sequencing, or to make some other fundamental change.

In connection with the above paragraph, I find that in spite of the fact that you collect

(continued on page 41)

(continued from page 3)

a profusion of figures in your company, there are not data on hand for either the problems of special causes or for measurement of the effect of common causes. Costly computers turning out volumes of records is not quality control. Figures fed back to a worker do more harm than good if they are devoid of signals that tell him (a) whether he himself is partly or wholly the cause of trouble discovered in product that passed through his operation, or (b) that the trouble arose from common (environmental) causes, beyond his control. The result is frustration and dissatisfaction of any conscientious worker. Without statistical signals, any attempt on his part to improve his work has the inevitable result of increases in variability and increases in costs.

Your production workers and your management need help that they are not getting. An important step, as I see it, would be for you to take a hard look at your production of figures — your so-called information system. Under more intelligent guidance, you would have far fewer figures but far better information about your processes and their capabilities, more uniformity, and greater output at reduced cost per unit.

I should mention also the costly fallacy held by many people in management that a statistician must know all about a process in order to work on it. All evidence is exactly the contrary. Competent men in every position, from top management to the humblest worker, know all that there is to know about their work except how to improve it. Help toward improvement can come only from outside knowledge.

Management too often supposes that they have solved their problems of quality (by which I mean economic manufacture of product that meets the demands of the market) by establishing a quality control department, and forgetting about it. In a sense, this is a good administration — to delegate responsibility and hold the man responsible to deliver the goods — but it is not working.

Why not? Most quality control departments work in narrow ranges of knowledge, with little concept or ability to understand the full meaning of quality control. Unfortunately, management never knows the difference. To grow up in a factory is not sufficient qualification for work in the statistical control of quality. There is no substitute for knowledge.

No good comes from changing the name of a quality control department to the department of operations research, or to systems analysis, or to some other fancy name.

Management too often turns over to a plant manager the problems of organization for quality. This man, dedicated to the company, wonders daily what his job is. Is it production or quality? He gets blamed for both. He is harassed daily by problems of sanitation, pollution, health, turnover, grievances. He is suspicious of someone from the outside, especially of a statistician, talking a new language, someone not raised in the manufacturing business. He has no time for foolishness. He expects authoritative pronouncements and quick results. He has difficulty to accustom himself to the unassuming, deliberate, scholarly approach of the statistician. The thought is horrifying to him, that he, the plant manager, is responsible for a certain amount of the trouble that plagues the plant, and that only he or someone higher up can make the necessary changes in the environment. He should, of course, undergo first of all a course of indoctrination at headquarters, with a chance to understand what quality control is and what his part in it will be.

Most men working in so-called quality control departments would welcome a chance to acquire more knowledge. One way is to send in a top-grade statistician on a regular basis for guidance. Another way is to send selected men in your company to one of the (few) statistical teaching centers, for two years. Your company needs desperately more statistical knowledge.

Statistical methods to improve training and supervision have not been utilized effectively in your company. Statistical evaluation of training and supervision, viewed as a system for improvement of skills and of operations, is an important part of quality control.

Perhaps the greatest problem (hardest to solve, I mean) is the perennially increasing shortage of competent statisticians that are interested in problems of industry. This shortage exists all over the world. Profound knowledge of statistical theory is necessary in quality control. Unfortunately, it takes around ten years beyond college, spent in study and internship under a master, to produce a competent statistician, and too few of the competent ones go into industry. This is partly the fault of management. A competent statistician will not stay in a place where he cannot work effectively and which fails to challenge his ability. The shortage of statisticians will continue. Meanwhile, companies must treat statistical knowledge as a rare and vital resource.

I find in my experience that management hardly ever provides organization and competent staff to carry on and develop control of quality on an economic scale. No one in quality control, however competent, can step in and work effectively in the absence of directive from the top. Proper organization and competence do not necessarily increase the budget for quality control. Management, in most instances, is already paying out enough money and more for proper organization and competence, but not getting their money's worth, getting tons of machine-sheets full of meaningless figures — getting rooked, I'd say, and blissfully at that. Your company is no exception.

I hold the conviction that here, as in Japan, it will be necessary for management to devote many hours to quality control, on a continuing basis, to learn something about the techniques, as management must hold themselves responsible for the problems of poor design, high costs, and quality, and must learn enough to judge the work of subordinates on these problems. No one is too important in a company, or paid too much money, to get some tutoring in statistical methods so that he can see better what the problems of the company are, and how his quality control people are doing.

W. Edwards Deming

Washington