On the Use of Judgment-Samples

W. EDWARDS DEMING, PH.D.
CONSULTANT IN STATISTICAL STUDIES

WASHINGTON 20016
4924 BUTTERWORTH PLACE

TEL. (202) EMERSON 3-8552

Reprinted from
Reports of Statistical Applications,
vol. 23, March 1976: pp. 25-31
(Union of Japanese Scientists and Engineers)

Rep. Stat. Appl. Res., JUSE.
Vol. 23, No. 1, March 1976

B-Section

On the Use of Judgment-Samples*

By W. Edwards DEMING
(Consultant in Statistical Studies)

1. Purpose of this paper

The purpose here is to try to answer some of the questions that statisticians have been asking for years about judgment-samples : (1) What type of problem requires, for best efficiency, use of a judgment-sample of blocks, plots, patients, clinics, hospitals, machines ? (2) What is the best allocation of the sample? Judgment-samples have been treated in a highly important article by Koller.** Elaboration of Koller's thesis will improve statistical practice, which is the aim of this paper.

The first step to an understanding of judgment-samples is to separate statistical surveys and experiments into two categories, enumerative problems and analytic problems. Once we understand the distinction between these two categories of problem, we may perceive at once answers to the questions posed. We shall see that for an analytic problem, where the aim is to provide a rational basis for action on a process, the results of a judgment-sample of blocks, patients, etc., with appropriate randomization of treatments, are amenable to the methods of conditional statistical inference ; but that for an enumerative problem, where action will be taken on the frame whence the sample was drawn, a judgment-sample may be subject to serious bias. We shall see also that optimum allocation of the sample is distinctly different in the two types of problem.

2. Enumerative studies and analytic studies, contrasted***

Enumerative : in which action will be taken on the material in the frame studied.

* Received Dec. 1, 1975. A paper delivered in Tokyo 28 November 1975 before the Japan Statistical Association.

** S. Koller, "Use of non-representative surveys for etiological problems", being Chapter 14 in the book edited by Johnson and Smith, *New Development in Survey-Sampling* (Wiley-Interscience, 1969): pp. 235-246.

*** 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 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. Having invented and used for years the terms enumerative and analytic, I continue to use them. My main concern here is the concept itself of the difference between the two kinds of study, regardless of what we call them.

The action to be taken on the frame depends purely on estimates of the numbers or proportions of certain categories therein. Put another way, the aim is description of the frame. How many farms or how many people belong to this or that category? The aim is not to find out why there are so many or so few people in a particular 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 small pox. 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. (The Bureau of Customs will calculate also from the samples how much duty to pay, if the ore comes from abroad.) 5. What is our share of the market for Product X in households of a certain characteristic ? 6. 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 (industrial product, wheat, patients, etc.) and will produce more in the future. The aim of a statistical study in an analytic problem is to try to learn what action on the process will bring improvement of product 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. What can we do to increase our share of the market ? 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 answers the question posed for an enumerative problem, subject of course to the limitations of the method of investigaton. 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. There is no such thing as a 100 per cent sample in an analytic problem.*

The two types of problem call for different procedures of selection and calculation. The theory of optimum allocation of effort to strata, stage-sampling, two-phase sampling, screening, estimation by use of regression estimators, are essential for economy in enumerative studies, and this is why most books on the subject are large and difficult. Thus, proportionate allocation of sample to strata, and Neyman allocation, with appropriate formulas for the calculation of estimates, modified by Hansen for different costs in strata, are examples of ways to extract the greatest amount of information per unit cost in an enumerative problem.

In an enumerative problem, we need not recognize strata at all in advance, nor even afterward. We get, without stratification, a valid estimate of a characteristic of the frame, though stratification might improve the precision obtained.

In contrast, in an analytic problem, it may be very important to investigate small strata as well as large strata to measure the differences between the two treatments *A* and *B*. It is usually best, in fact, to investigate at the start strata at the expected extremes of response. Failure to perceive in advance which strata may react differently to *A* and *B* may delay badly the results of an analytic study. Optimum allocation of the sample to strata for an analytic problem is almost always for practical purposes simple equality

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

($n_1 = n_2$). Modification for variances and costs that are widely different is rare but simple, namely, take n_i proportional to $\sigma_i/\sqrt{c_i}$.

3. 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, we should not 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. Statistical theory enables us to minimize the net loss from paying too much for the material and from doing too much testing.* ** ***

4. What do we need in an analytic problem ?

What we try to learn in an analytic study is whether the difference between the two treatments A and B appears to be of material importance, economic or scientific. This required difference we designate by D. Symbolically,

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 in the future ?

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

If it appears pretty definitely from the experiment that B is superior to A by the amount D, in some stratum, then the experiment has discovered something. The next question is, what other strata will show a similar difference, or the reverse ? An analytic study proceeds sequentially.

5. Two kinds of error in an analytic problem.

The question in an analytic problem is what action to take on the process : What to do ? How will the two treatments compare in the future ? We may make either one of two mistakes :

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

These two mistakes differ fundamentally from the two mistakes that we can make in

* Leo Törnqvist, "An attempt to analyze the problem of an economical production of statistical data", *Nordisk Tidsskrift for Teknish φkonomi*, vol. 37, 1948: pp. 263-274.

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

*** Robert Schlaifer, *Probability and Statistics for Business Decisions* (McGraw-Hill, 1959), Sec. 33.3. I. Richard Savage, *Statistics: Uncertainty and Behavior* (Houghton Mifflin, 1968), Ch. 4.

an enumerative problem.

The function of statistical practice is to try to minimize the net loss from both kinds of mistakes, whether the problem be enumerative or analytic. Analytic problems unfortunately present limitations and challenges in inference that are not often recognized.

6. Limitations of statistical inference

All results, whether the problem be enumerative or analytic, 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 investigator, who carries out the interviews or measurements. In addition, (d) the results of an analytic study are conditional also on certain environmental conditions, forced or prescribed, such as the geographic location of the comparison, the date and duration of the test, the soil, the climate, description and medical histories of the patients or subjects that took part in the test, the observers, the hospital or hospitals, the 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 wide range of conditions, may differ greatly under other conditions—other soils, other climates, 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 shows up well in a test that covers three weeks may cause grief and regret after a few months. A drug that at first shows great promise may later turn out to be a disappointment because of unforeseen side-effects.

Accordingly, in an analytic problem, one can only formulate a loss-function on a conditional basis in an attempt to minimize the net economic loss from the two mistakes, conditional on the environment of experiments conducted in the past. Tests of hypotheses and significant differences are not the answer.* Experimentation must be supplemented by knowledge of the subject-matter (chemistry, pathology, psychology, engineering, medicine, agricultural science). The statistician renders his best service by understanding and explaining the limitations of the conclusions that can be drawn from a study.

7. Use of judgment-samples

We are now ready to try to answer the questions posed at the outset of this paper.

The result of a judgment-sample for an enumerative problem (e. g., to estimate the frequency of a certain type of error in a frame of accounts receivable) is worth no more than the reputation of the man that signs the report. The reason is that the margin of uncertainty in the estimate reported is entirely dependent on his knowledge and judgment. Expert opinion on the margin of uncertainty in use of a judgment-sample for an enumerative purpose may of course be useful; there are rare circumstances where a probability-sample is extremely difficult and costly.

---

* W. Edwards Deming, "On probability as a basis for action," *American Statistician*, vol. 29, No. 4, 1975: pp. 146-152.

Fortunately, 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 ship-load of ore where one can only take samples from exposed portions. There are usually ways around such difficulties, namely, to draw samples while the coal or ore is being loaded or unloaded.

In contrast, one may say that for most analytic studies we have never had anything but judgment-samples. Most of man's knowledge in science has been learned through use of judgment-samples and careful inference. 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. The climate, rainfall, soil, and other conditions during a season at Rothamsted constitute a process. We can by experimentation obtain an $\bar{x}$ for any variety (the average yield per acre over a set of trials), under a given set of conditions. It is futile to speculate on the existence of $E\bar{x}$, as this process can not be repeated, not even at Rothamsted : climatic conditions will never be the same there again. Certainly they will be different in some other part of the world, where we need to apply the results. Even if $E\bar{x}_A$ and $E\bar{x}_B$ for varieties A and B existed at Rothamsted, we could not assert on the basis of statistical inference alone from an experiment at Rothamsted that the difference between A and B obtained there would tell us anything about the difference between A and B in our areas, where we shall raise wheat next year.

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 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.

8. Importance of design of experiment

We appreciate now the importance of theories of experimental design and of theories of inference, in order to build into an experiment the greatest possible efficiency, to make it as productive as possible. But every inference (conclusion) is conditional, no matter how efficient be the design of the experiment. A conditional inference does not permit generalization : we can not assert, on the basis of a conditional statistical inference alone that other patients, other hospitals, other pupils, other locations, would show similar differences nor greater differences. We fill in the gap by knowledge of the subject-matter.

Conditional inferences may be drawn by methods of estimation, which include paper and pencil plots, correlation diagrams, run-charts. Tests of significance fail to provide a basis for action.*

---

* W. Edwards Deming, 1975; *loc. cit.*

9. Optimum allocation of effort in experimentation

The fact is that statistical workers in experimental design, in the early stages of comparison of two treatments, or of two mechanisms, have an easy road compared with the statistical worker in enumerative studies. This follows from an earlier paragraph : when the experimenter discovers by any means any stratum (areas, hospitals, patients, mechanical designs) that under recognizable conditions yields with high probability a substantial difference D between two treatments, or yields dependably a difference of no consequence, he has made a contribution to knowledge.

This contribution to knowledge will usually be incomplete, but it is nevertheless a contribution.

The statistical worker in experimental design may thus bite off strata one at a time, as results seem to indicate, until he is able, in his substantive judgment, to establish a classification of areas and conditions under which the superiority of B over A by the amount D is established, or is inconsequential.

Not so with enumerative studies. Omission from the frame of a stratum of substantial size and of unknown characteristics (e. g., the people that fail to respond) may defeat the aim of the study (total revenue, total dollar-value, sales of a certain kind of item, prevalence of a specified affliction), or require re-definition of the universe of study.

In the early stages of an analytic investigation, it is nearly always the best advice to start with strata near the extremes of the spectrum of possible disparity in response as judged by the expert in the subject-matter, even if these strata are rare. In illustration, I may mention that only last week, a manufacturer of clinical thermometers with a claim to superiority in an important characteristic (speed of reaching equilibrium), proposed to conduct comparisons on patients with normal temperatures (people in his own office would do) and on patients with high fevers. This would be a good beginning, I thought, though he must first settle on a statistical criterion for the characteristic to be compared. If the superior performance is not obvious at either extreme (normal temperature or fever), then he might well question not only the alleged superiority, but also his definition of what is superior.

An auditor may use a probability sample to estimate the total net dollars receivable in a frame of 150,000 accounts. But he has an obligation to investigate separately any suspected source of error. For a suspected stratum, he may draw a sample of accounts with equal probabilities, or he may prefer to study in particular all the accounts of some specific type.

Tests of some component part in an automobile such as power brakes or steering, should be conducted at extremes of possible stress, and beyond.

If one were to try to measure the difference in cost of handling weights of 100 pounds and weights of 400 pounds on platforms of carriers of motor freight, one would, I believe, make studies on a varied selection from platforms (a) that are fully mechanized, and from platforms (b) that are only partially mechanized, and from platforms (c) that are not mechanized at all. I would use random numbers and replicate the design, but I would not give equal probabilities to all platforms in the country.

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 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 example comes from Koller (*loc. cit.*, p. 237):

> 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.

And on page 246, he adds, in effect .

> For an etiological survey, tests in all areas of a population may be the wrong way to proceed.