ON SOME CRITICAL POINTS IN SURVEY SAMPLING *

By

W. Edwards Deming
Consultant in Statistical Studies
Washington 20016

For criticism, not to publish

Scope of this paper. First, what is sampling? Interpretation of an observation is a problem of sampling. In professional statistical practice we make sure in advance that the observations planned will be useful in the judgment of the man that will pay the bill. Sampling is the science and art of planning the observations and of making calculations based on them and on other information to be as helpful as possible, within an allowable budget, to the man that will use the results.

This paper will mention some of the practical problems of sampling that are not in the books. My purpose in writing the paper is to try to contribute something to the improvement of statistical practice. Treatment of the various topics will be lop-sided--heavy on topics that are generally misunderstood, light on others, complete omission of some.

A course in sampling is not the aim here. One can learn the theory from the right books. All that it requires is to devote several years to self-study or to classes. The next step is to work 3 to 5 years under a master. The material and administrative restrictions in practice change from one problem to another. To understand what theory to use and how to adapt it to a problem, to be able to create new theory as required, and how to explain the results of a study in a way that will be helpful to the man that will use the results is not in the books. One reason is that these things are hard to write about. Success in statistical work is theory plus creative art, and the ability to understand the problems of the man that will try to use the results.

The first question to ask, when a statistical investigation is proposed, is this: can you get from any unit in the frame (person, household, farm, business establishment, piece of product), no matter how you select it, the information that you require? If the answer is yes, then ask a second question: if you were to elicit the desired information from every unit in the frame, would the compilation of results be useful? If the answer is yes to both questions, then a statistical investigation might be worth consideration. If the answer to either question is no, then take a second look: revise the aims, or the method, or abandon the survey.

The difference between a statistical investigation and one not statistical is that with the statistical investigation, we are able, in the end, to evaluate the main sources of uncertainty in the results. The more we know about what is wrong with a figure, the more useful it becomes (John Tukey; or for illustration, almost any paper on experimental physics or chemistry). We often learn more from what went wrong in an experiment than from what went right (Paul Olmstead). Ability to evaluate the uncertainties

* Prepared for presentation and criticism for the Princeton Conference on Applied Statistics sponsored by the Metropolitan Section of the ASQC and the Princeton Biopharmaceutical Subsection of the American Statistical Association. Revised for the National Conference of the Statistical Reporting Service, Washington, 18 February 1975.

in a result is not an accident. It is possible only by use of appropriate statistical design, which will include statistical controls for the detection of blemishes and blunders in the measurements, interviews, and processing, to be in position to evaluate the possible effects on the results that arise from the chief sources of nonsampling errors.

The theory of sampling, besides uses in demographic, social, and medical surveys, marketing research, the statistical control of quality, accounting, has application in many other ways. Thus, the theory of failure, average life of complex equipment, theory of reliability, design of experiment, optimum policy of maintenance, queueing theory, optimum policy of inventory, are all applications of the same body of theory. The statistical control of quality embraces application of all known statistical theory.

Probability methods of sampling are now required for statistical evidence presented to the Interstate Commerce Commission and to the Federal Communications Commission; perhaps also to other agencies.

Meanwhile, in contrast, some government agencies and most of industry and most marketing research companies muddle along. I could, from my own observations, write a chapter of horrors about disappointments and financial loss that could only be blamed on to results derived from bad samples. Companies that have competent statisticians usually fail to make use of them company-wide. The world-shortage of food that we read about in every newspaper these days is attributed by Milton Godfrey * to failure at top levels of government in several countries to understand the rudiments of optimum policy of inventory-management. There are certain to be fat years followed by lean years. Not every year comes up to the average. To cut the acreage of wheat following a surplus is to reinforce the amplitude of the natural variation of climate and environmental factors.

Statistics are a basis for action. Any investigation is for purposes of action. To acquire new knowledge is to take action to modify existing knowledge. One can not simply store new knowledge, to retrieve it later.

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 treatments in medicine and in agriculture, improvement in quality of product, break-down of service for purchased equipment and for rented 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.

These problems can not be understood and can not even be stated, nor can the effect of any solution be evaluated, without the aid of statistical theory and methods. The only knowledge that has meaning, whether for business or for legal purposes, is statistical. One can not even define operationally 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, for example), to have

* Milton Godfrey, "Future challenges," address to the New York Chapter of the American Statistical Association, 14 Nov. 1974.

any benefit to mankind, or in fact to have legal force, must specify the capability of the performance of prototypes or of items in actual service. Performance and capability can be defined only in statistical terms. A standard that can not be defined in statistical terms is no standard at all.

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?

What is due care in manufacture? The only possible answer is in statistical terms. The most that a manufacturer can put into the uniformity and dependability of a device is (a) to achieve and maintain statistical control of the most important quality-characteristic of the main components and ingredients, and (b) to be able to demonstrate by adequate statistical records the results of his tests and controls, along with action taken on special causes and on common causes, and statistical evidence of the results. Engineering judgment, not statistical theory, must specify what quality-characteristics are important.

In spite of scrupulous care and intelligent use of statistical controls, it is inevitable that a defective item will get out now and then. An unfortunate freak of this kind can not be viewed as prima facia evidence of carelessness. Punitive action should be directed against failure of the process and against lack of statistical evidence in regard to it, not against the freakish defective.

The manufacturer will of course stand behind any warranty for strict liability. Does a physician practice under conditions of strict liability? How could he? What about statistical practice?

The distinction between enumerative and analytic studies (vide infra), if explained patiently by statisticians to the legal profession in industry and in government, would clear up many problems about safety and reliability, eventually, I dare say, even mal-practice in medicine.

Statistical inference, to be effective, must interpret results as a basis for action. Drastic changes in practice and in writing and in teaching are called for. As Shewhart said, the standard of knowledge and the requirements of workmanship in industry and in public service are more severe than the knowledge required in pure science. He ought to have added teaching along with pure science.

It is time that statisticians became involved with problems, even if the knowledge and fortitude required for statistical practice are more exacting than for teaching.

The frame, the universe, environmental conditions. A statistical study proceeds by investigation of the material in a frame.* The frame is an aggregate of 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, manufactured parts, or other identifiable units that would yield useful results if the whole content were investigated. Equally important in an analytic problem (vide infra) is the specification of the environmental conditions such as the season, climates, rainfall, levels of light, radiation, fertilizer, ranges of concentration, dosages, duration of the test, pressures, temperatures, speeds, voltages, or other stresses that the material or product will be subjected to in the experiment.

Steps in the design of a study. The following steps refer to studies in consumer research. The reader may adapt them to his own field.

1. Satisfactory re-statement of a problem, and formulation thereof in meaningful terms (statistical model), so that a statistical investigation will yield as much useful information as reasonably possible within allowable costs and within the framework of skills and of other resources that are available.

2. Specification and preparation of a suitable frame for sampling (lists of areas, business establishments, people, items, or other units, range of temperature, concentration, speed, etc.).

3. Statement of the ranges of climate, soils, dates, stress, length of time, that the study is to cover.

4. Specification of the procedure for the selection of a suitable number of sampling units for investigation (invariably with random numbers; sometimes with stratification, sometimes with complex sampling units and fractional probabilities). The sampling procedures contain rules for recalls on people that were not at home on previous attempts.

5. The formula or procedure for the formation of estimates, appropriate to the procedure of selection. The central problem in the theory of sampling is the design of procedures for selection and estimation so as to maximize the amount of information per unit cost, within any restrictions imposed.

6. The formula or procedure for the formation of standard errors, to measure the margin of uncertainty that arises from variability in the material and from variable performance of the investigators, coders, punchers, and other workers.

7. Design statistical controls (more later)

-- to assist supervision by detection of non-uniform performance in the field and in the office.

---

* First defined, without use of any specific term, by F. F. Stephan, American Sociological Review, vol. 1, 1936: pp. 569-580.

-- to detect and evaluate nonsampling errors, especially persistent operational blemishes of investigators, coders, punchers, and other workers.

After the survey is finished, comes the interpretation of the results. This will be a report intelligible to the user of the data. Upper or lower fiducial limits, or both limits (depending on the problem), and their interpretation may be important, provided the execution of the survey did not depart too far from specifications. It is equally important to take into account possible uncertainties that could arise from persistent or large accidental operational blemishes detected in the statistical controls, or otherwise. The business of fiducial limits is statistical; likewise the statistical plan for the detection of operational blemishes in the statistical controls. However, the evaluation of blemishes so detected is not entirely statistical, but must depend heavily on knowledge of the subject-matter. For example, the nonresponse in a survey was 20%; what possible biases may one attribute to nonresponse? The statistician can be of inestimable help to the expert in subject-matter by shielding him against unwarranted conclusions.

Interpretation of statistical tests is one of the most difficult parts of statistical work, and the methods described in statistical books under estimation and tests of significance are woefully inadequate. Again, pencil and paper are the chief tools of analysis.

Type I. Structural limitations, or built-in deficiencies of the questionnaire or method of measurement

1. Failure to perceive in advance what information would be useful. Failure to elicit information that, in the end, turns out to be needed badly.

2. Inept wording or sequences of questions.

3. Inept choice of frame (area, list, hospital, manufacturing plant).

4. Unfortunate choice of date or of other environmental conditions for the survey.

5. Ineffective rules for coding.

6. Ineffective tabulations, such as classifications and class intervals not well suited to the consumer's needs.

7. Bias arising from wrong weighting, or from incorrect adjustment.

Type II. Operational blemishes and blunders

8. Small errors of a non-cancelling nature (e.g., omission of a sampling unit; instrument out of adjustment).

9. Persistent favor of an investigator toward one direction, causing an operational bias.

10. Large errors, such as a single-time blunder, reporting a final result as 86.8 in place of 68.8 (it happened).

11. Drift of instruments or of observers.

Type III. Random variation

12. Random variation arises from differences that exist by anybody's standards between units, whether they be accounts, households, blocks, segments, manufactured items, and from small accidental independent variations of observation, which vary with the time of day, direction of travel, etc. The standard error is a universal measure of random variation.

Type IV. Unwarranted inferences on the part of the user

13. Failure to understand the conditional nature of the results of an experiment (to be elaborated), being mis-led by statistical calculations.

Statistical controls for detection of nonsampling errors. Some people suppose that once they write a set of instructions, and rehearse them with the people that will carry out the work, all will go off well, just as planned. This is a costly fallacy. It is better to monitor the work at random points and at random times, to observe what is happening, to straighten out misunderstandings, bear down on laxity, and (as frequently happens) to change procedures where a better way comes to light.

In my own work, I find it possible to introduce "control points," for re-interviews or re-inspections of somewhere between 1 in 6, 1 in 10, or 1 in 20 of all units investigated in the main sample. It is a simple though tedious matter to lay the sample out to designate in advance which investigator will re-test a selected unit.

The re-tests are not used in the tabulations but to detect troubles before they become serious and to know in the end the effects of failures of various kinds.

Statistical controls are not a luxury: they are simply good management. In a recent experience with Canadian Facts, Ltd., in Toronto, it was possible through skilful arrangement for immediate hand tabulation, to detect high variance between interviewers on certain questions the first half day on the job, before it was too late to take corrective action in the form of revision of the questions and improved methods of interviewing and supervision.

Statistical controls in the Census under the guidance of Morris H. Hansen reduced the costs of interviewing in the Census to a third of the cost of interviews carried out in commercial research, while continually increasing the accuracy of results.

A quarter of the total budget invested in controls well over-pay their cost. Random re-interviews, studies of records of time spent on travel and on interviewing, tests of the questionnaire and of the coding and processing, with corrective action bring improvement in the questionnaires, coding, training, supervision, mode of travel, time of day of interviewing.

Human error in studies is only one source of nonsampling error. Illegible entries, wrong extensions, wrong counts, cause their share of trouble. An ever-present difficulty in my own work is the widespread failure of manufacturers to keep their instruments in good condition. Interpretation of figures derived from an investigation is often cloudy for this reason.

Results of an investigation of any study can not be better than the workmanship that goes into the investigation, nor better than the instruments used. The word sampling repeated 100 times is no guarantee of accuracy. It is sometimes stated that a sample is better than a complete census. One should say that a small sample has the possibility of being better than a complete census, as the investigators in a small sample can be more carefully selected, trained, and supervised, and instruments kept in better condition.

Differences between investigators. It has long been known that an interviewer injects a certain amount of influence into answers, especially if she is not continually tested and retrained on the basis of statistical tests. Thus, the interviewer will introduce causes of similarity amongst the interviews that she conducts. Many examples have been cited in the literature.

Some people talk about facts, as if there were a correct value for the number of people that ever heard of Brand X of oven-cleaner, or that bought it during the past three months. One might suppose that there is a definite number that would be obtained for the count of people unemployed, were a census taken, or for the number of families that need welfare.

The truth is that all that we ever have is the results of interviewing and coding. If you change interviewers, or change only 15% of them, on a new survey, you may expect to get different results even from the same households. History itself is only what somebody thought was worth while to record, and how he recorded it.

These dreary truths do not mean that no data are any good, nor that the situation is hopeless, nor that there is no use to study history. Rather, they mean that, to use data, one should understand the process that produces them. The figures that one derives from a survey are the end-product of a chain of operations.

In the days when differences between investigators was not appreciated, the problem never came up. One may recall a quotation from E. C. Fieller:

> Before the inherent variability of the test-animals was appreciated, assays were sometimes carried out on as few as three rabbits: as one pharmacologist put it, those were the happy days.--Supplement, Journal Royal Statistical Society, vol. vii, 1940-41: p. 3.

In a study of diagnosis and change in condition of mentally disturbed people in a community, I learned that a report on a patient (improved, about the same, worsening condition, deterioration, etc.) was recorded as a consensus of two physicians. Now anybody knows that a consensus is likely to be a recessive gene swallowed up by a dominant gene.

This was a severe loss of information and a loss of a basis for training. It would be better to let each physician record independently his opinion, with brief notes on his reasons; then to compare notes. The comparison of opinions and reasons for difference would provide a meaningful basis for continual training and improvement of the physicians. Without knowledge of the variance between physicians, the question of whether a treatment is effective remains clouded. There may be more variance between patients within physicians than between physicians within patients.

The two physicians should of course argue about some cases where they agree with each other. To conduct only investigations where they disagree is biased, as we know from work of W. J. Youden on the bias of the chemist's rule to adopt the best two out of three observations.

The physician in charge adopted the suggestion and is using his influence to spread the idea.

Knowledge of the variance between investigators enables one to fix the economical size of sample. Thus, in the circumstance that interviewers have been trained so well that, for important characteristics, the intraclass correlation between sampling units within interviewers is as low as .03, a complete census of an area is no better than a 25% sample of households that are handed a self-filled questionnaire to mail in. This may be seen from the equation

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}[1 + (g-1)\rho]$$

in which we place $\rho = .03$, $g = 100$ households for the average work-load of an interviewer. The bracket then takes the value $1 + 3 = 4$, which indicates that the variance of $\bar{x}$ is equal to what it would be with a 25% sample and zero door-to-door correlation. n is the number of households in the sample.

This equation, or rather the understanding of it and use thereof have saved the Census millions of dollars.

Savings and improved reliability in the same proportion are possible in research of many kinds, in industry and in government.

Every manufacturer would like to know how many people ever heard of his product. Unfortunately, the question, "Have you ever heard of Brand X?" is subject to such high intraclass correlation between respondents within interviewers that the effective size of sample is little bigger than the number of interviews (see table).

Possible magnitudes of the loss of information from variances between interviewers is illustrated by the accompanying table, which is arrayed by F-value, from lowest to highest. F is defined here as the total variance

Effect of variance between interviewers

	Question (asked of female home-makers)	$\hat{p}$ Proportion Yes	$\hat{\sigma}_{\hat{p}}$	F
1	What brand of canned or packaged coconut did you buy most during the past 12 months? Proportion that answered Brand A. (Base: people that had bought any canned or packaged coconut.)	.029	.003	1.29
2	Where did you last buy a rinse, tint, tone, or lightener? Base: women that work on their own hair at home, never at a beauty shop. Proportion that answered at a drug store.	.590	.082	1.45
3	Have you ever used canned or packaged coconut? Proportion yes.	.733	.010	1.46
4	Of those that had purchased an aerosol medicated room vaporizer-decongestant, the proportion that say that they are very well satisfied.	.862	.063	1.57
5	Age of head of household with a female homemaker. Proportion in age-group 25-29.	.086	.008	1.64
6	About the country that you would least wish to visit, is it exciting, interesting? Proportion yes.	.047	.006	1.76
7	Proportion that had ever heard of a specific brand name for a gadget on a washing-machine.	.652	.041	1.81
8	I have here three brand names. As I read each name, tell me whether you have ever heard of it. Proportion that had heard of Brand X of analgesic	.040	.006	1.86
9	Income of whole household, proportion between $3500 and $3999 annually.	.037	.006	1.88
10	Of people that use aerosol furniture polish or wax, how many cans did you buy last year? Proportion that bought 2 or more cans.	.825	.015	1.89
11	Marital status. Proportion single.	.029	.006	2.00

	Question	$\hat{p}$ Proportion Yes	$\hat{\sigma}_{\hat{p}}$	F
12	Have you ever heard of Brand F of oven-cleaner? Proportion yes.	.103	.011	2.21
13	Of the people that had ever heard of a special arrangement of tubs in a washing-machine, what company makes it? Proportion that said that Company A makes it.	.234	.016	2.24
14	Have you ever heard of Brand G of oven-cleaner? Proportion Yes.	.463	.019	2.26
15	Highest grade of school completed by head of household. Proportion that had completed the 8th grade.	.130	.109	2.31
16	The respondent is shown an actual jar of a certain deodorant. Do you recall seeing it before? Proportion yes.	.830	.016	2.56
17	Proportion of females that are other females (not female home-makers) in households with a female homemaker.	.173	.016	2.84
18	Have you ever heard of Brand H of analgesic? Proportion yes.	.696	.019	3.04
19	Proportion of households with a female home-maker and with 2 or more people in the house-hold.	.912	.015	3.51
20	Proportion of households in which the husband of the female homemaker wishes to go camping, more than on any other vacation..	.234	.031	5.01
21	Does Food Company X make Brand Q of breakfast cereal? Proportion Yes.	.212	.028	5.11

between interviewers divided by the binomial variance. A definite pattern seems to emerge. Questions that beget a waivering answer lead to high F-values, and to little useful information. The high F-value for the proportion of women that are single might be taken to indicate that some women are not sure of the answer. A better explanation may be that the interviewer did not ask the question, but supplied the answer herself.

The remedy seems so obvious. An esteemed friend, in charge of marketing research for a large corporation, in a discussion after I had presented figures on variances between interviewers, came through with the adorable thought that all these problems would disappear if one would only standardize the procedures so that all interviewers performed equally! He had never thought it necessary to spend a dollar of his huge budget to learn how the interviewers were performing in the work that he paid for. Some people prefer to remain happy.

Replicated designs for simplicity in computation of variances. A plan that I follow in all kinds of work, whether it be accounting, inspection of physical equipment (telephone plant), marketing research, surveys of communities, traffic, interviewing by telephone, medical studies, or other, is this:

1. Lay out the whole job in subsamples, each subsample being a valid sample of the frame. The subsamples may be random selections within strata, or systematic with random starts at every 8th zone (Mahalanobis; Tukey).

2. Allot a random portion of each subsample to each investigator.

On this design, every interviewer has a valid sample of every subsample, hence a valid sample of the whole sample. Each subsample is affected equally by the variance between interviewers, except for improbable interactions.

Tabulations by subsample and by interviewer facilitate for any characteristic tabulated by subsample and interviewer, immediate estimation of the standard error and of the total variance, including the variance between interviewers. * **

The number of degrees of freedom for the variance between subsamples may be any number from the number of subsamples on up by tabulating a number of sub-totals.

If the total variance far outweighs the variance between the subsamples, owing to high variance between interviewers, then the standard error should be based entirely on the total variance. The degrees of freedom will be the number of interviewers diminished by unity. This is a severe loss in the power of analysis, which fact emphasizes the need of statistical controls in the training before it is too late.

---

* P. C. Mahalanobis, "On large-scale sample surveys," Phil. Trans. Royal Society., vol. 231B, 1944: pp. 329-451; "Recent experiments in statistical sampling in the Indian Statistical Institute," Journal of the Royal Statistical Society, vol. cix, 1946: pp. 325-78.

** W. Edwards Deming, SAMPLE DESIGN IN BUSINESS RESEARCH (Wiley, 1960), Part II. W. Edwards Deming and Morris H. Hansen, "Some theory on the influence of the inspector and environmental conditions, with an example," reprinted from Statistica Neerlandica, vol. 26/3, 1972: pp. 101-112.

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.*

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 counts of the frame. The aim of a statistical study in an analytic 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? 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 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.) 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 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 answers 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. There is no such thing as a 100 per cent sample in an analytic problem.**

---

* 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 shall continue to use them, for a while at least. My main concern here is the concept itself of the difference between the two kinds of study, regardless of what we call them.

** 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.

The two types of problem call for different procedures of selection and calculation. Optimum allocation of effort for estimation of the total number of units in a frame that have a prescribed characteristic, or for the average per unit, as called for in an enumerative problem, is treated extensively in all textbooks.

Optimum allocation of the sample to strata for an analytic problem is almost always for practical purposes simple equality ($n_1 = n_2$). Modification for variances and costs that are widely different is rare but simple, namely n_i proportional to $\sigma_i / \sqrt{c_i}$.

I venture the opinion that controversies that have been witnessed between statisticians may be viewed as one man working on enumerative or descriptive problems, the other man working on analytic or comparative problems, both right in his own sphere, but neither one aware of the existence of the other type of problem.

In an enumerative problem, we need not recognize strata at all in advance, nor even afterward. We can get a valid estimate without stratification, though stratification might improve the precision obtained. Then too, there are ways to use a preliminary sample and to stratify only it, not the whole frame.

In contrast, in an analytic problem, it may be very important to recognize in advance the existence of small strata and to investigate them as well as large strata for differences between the two treatments A and B. Failure to perceive in advance which strata may react differently to A and B may greatly impair an analytic study.

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. Without use of statistical theory, there is no way to minimize the net loss (paying too much for the material offset by the cost of testing). * ** ***

---

* Leo Törnqvist, "An attempt to analyze the problem of an economical production of statistical data," Nordisk Tidsskrift for Teknish Økonomi, 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.

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. The statistician renders his best service by understanding and explaining the limitation of the results of a study (next section).

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 can not acquire enough empirical data to predict performance under the environmental conditions of the future.

For illustration I take a recent case wherein a group of carriers of motor-freight sought higher rates for hauls under 400 miles. The evidence consisted of an exhibit of operating-ratios (cost/revenue) covering hauls in 1973 of under 400 miles and covering hauls of 400 miles or over. The data came from a continuing sample of about 50,000 freight bills from 89 carriers. Statistical controls applied over a 10-year period have eliminated the main sources of uncertainty from operational blemishes and blunders, different interpretations of definitions of line-haul and kindred troubles.

There could be no doubt that the average operating-ratios of hauls below and above 400 miles are different. We know this in advance without making a study. The big question in the minds of the carriers, and of their opponents (mainly short-haul shippers), and in the minds of the Interstate Commerce Commission, was whether the operating-ratios for long and short hauls will in the future be enough different to warrant different rates. An excerpt of the results is shown in the inserted table. The number of degrees of freedom in the estimates of standard errors is in the hundreds.

1973

Weight and mileage		Operating-ratio	Standard error
Under 500 lbs.	miles 0- 400	137.8%	5.9%
	401-1050	118.1	1.2
Difference		19.7	6.0

It had been decided earlier by the carriers that a difference of 3% = D in operating-ratios for long and short hauls is big enough to warrant action on the matter of rates. The estimated difference between the operating-ratios for 1973 was 19.7%, which was more than 3 times its standard error. The conclusion was inescapable, based on the table inserted here and on other evidence not produced here, that the difference in operating-ratios for 1973 exceeded the critical D = 3% and accordingly warranted action on the matter of rates.

Action on the rates will apply to future years. The carriers can make either of two mistakes: raise their charges too much for mileages under 400, and regret it later, or not enough, and regret it later. Future years will be different from the past--higher costs of fuel, different type of contract with the drivers, tendency toward consolidation of small shipments to get the benefit of freight forwarding or truckload rates or carload rates, and thus defeat the increase. More data for the past year could not reduce in any important way the risk of either mistake.

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, the results of an analytic study are conditional also on certain environmental conditions, proscribed 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 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?

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. (see next section). 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 cover three weeks may cause grief and regret after a few months.

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 extend the coverage. Only the statistician, with his knowledge of theory, can perceive the limitations of the results that will come out of a study.

My favorite question when someone proposes a study 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.

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 consequence.

Likewise, tests of whether the data of a survey or an experiment fit some particular curve is of no scientific or economic importance. $P(\chi^2) \to 0$ 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 whether it is helpful in the discovery of the cause of some observed anamoly, or whether some other curve would do better. How robust are the conclusions?

Statistical significance by itself thus conveys no knowledge, no basis for action.

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. Why speak of testing the hypothesis that A = B? Statisticians can put their knowledge to more useful endeavours.

The fallacy of tests of hypotheses has been exposed here and in many places. A further difficulty is that the functions used in the analysis of variance, and in estimation, are symmetric. Interchange of any two observations x_i and x_j leaves unchanged the mean, the variance, and even the distribution itself. In other words, use of variance buries the information on order of appearance in the original data.

The presentation of results, to be optimally useful, and to be good science, must conform to Shewhart's * rule to the effect that the data should be presented in a way that will preserve all the evidence in the original data for the uses intended.

Pencil and paper for construction of distributions, scatter diagrams, run charts, comparison of results from small groups, are more efficient than analysis of variance, preserving the information in the original data. In fortunate circumstances (normal estimates, independence, absence of patterns), and when the whole study went off just as intended, one may indeed summarize the results of comparisons by use of estimation or by likelihood. But these circumstances require demonstration by simple methods of pencil and paper.

* Walter A. Shewhart, Statistical Method from the Viewpoint of Quality Control (Graduate School, Department of Agriculture, Washington, 1939) p. 88.

What do we need? What we need to know is whether the difference between A and B appears to be of material importance, economic or scientific. This required difference we designate by D. Symbolically,

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

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.

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 experiment may of course discover the converse. It will in any case indicate better ways for the design of the next study, the aim of which will be to remove some of the areas of doubt.

Criticisms of teaching. I 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 latter 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 on a job as a statistician were to make a calculation like this, and a statement like this, 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? It is well to give students a chance to think and to ask questions of the data, but how could they? The original data are not in the text.

Are students made aware that the standard errors 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, were 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. But who cares? What difference would be 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). Define H_0 as the hypothesis that $\mu_1 = \mu_0$; H_1 the hypothesis that $\mu_1 \neq \mu_0$.

> 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).

Comment. Small wonder that students have trouble. So do I.

Why? When I enquire of authors and teachers, to learn why they persist in teaching illustrations like these, the answers follow a pattern like this.

1. The theory is expressible in formulas.
2. We can teach formulas.
3. Students can learn the theory and take an examination on it.
4. If they understand the theory, they can teach it.
5. Editors require authors to include tests of significance: otherwise an author's paper will be rejected.
6. The departments of psychology and education require us to teach it.

---

7. A student will not mis-use theory later in practice if he truly understands it. Besides, many of our students will teach statistics, and not be in practice.

Comment. We have little assurance that students will not later, in practice, mis-use the theory that he learned in school. Even if that were so, which it is not, why, I ask, derail the student at the start?

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, true value, and level of significance.

A teacher that gets involved in problems will have a basis for making a choice of what theory to teach, and he will have illustrations of his own for the classroom and for his books. He will not have to borrow from other books.

Involvement in a problem carries responsibilities. The statistician in practice must write a report on the statistical reliability of the results, what they may mean and what they don't mean. This report states the possible margins of uncertainty from sampling variation and from operational blemishes big and little discovered in the controls, the nonresponse, illegible 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, the questionnaire if one was used, and the difficulties encountered. Involvement in a problem means the possibility of facing a board of directors, or facing cross-examination.

The statistician in practice works with experts in subject-matter that are sometimes skeptical about statistical methods (sometimes with reason). The expert in the subject-matter may have had a course or two or three in statistics, and may thus know a lot that is not so. His position is the reverse of what it is in the classroom, where almost anything will do. The easy way out is not to get involved: just give advice now and then and leave it to someone else to follow up on it. Mere desultory advice in the department of double integrals does not constitute involvement in a problem.

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

1. Use of a judgment-sample (e.g., by which to estimate the frequency of error of a certain type in a frame 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 offer relief from the uncertainty of the magnitude of sampling variation in a result. They also put blinders on the problems of nonresponse, and they eliminate occasional high cost and inconvenience of travel, but not the biases of such elimination.

* I am deeply indebted to my friend and colleague Professor S. Koller of the University of Mainz for the need of emphasis on the importance of judgment-samples in analytic (comparative) studies.

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 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.

3. In contrast, one may say that for most analytic studies we have never had anything but samples of convenience. Most of man's knowledge in science has been learned through use of judgment-samples. 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 $E\bar{x}_A$ and $E\bar{x}_B$ at Rothamsted would be related to $E\bar{x}_A - E\bar{x}_B$ in our area, where we shall raise wheat next year, and must decide between A and B.

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 chosen for convenience may constitute important new knowledge.

Randomization within the blocks in a convenient area (for trials of wheat), or of patients (for comparison of treatments), removes an important area of doubt, and justifies the use of probability for conditional inferences.

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. Further experimentation would be required unless the gap can be filled in by knowledge of the subject-matter (medicine, chemistry, engineering, psychology, agricultural science, etc.).

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 the preceding paragraph: when the experimenter discovers any stratum of a judgment-selection of materials (areas, patients), and under any conditions, a substantial difference D between two treatments, he has made a contribution to knowledge, whether he ever carries out another experiment or not. A difference that turns out to be inconsequential is likewise 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 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 that 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.

Not so with enumerative studies. Omission of a stratum causes a deficiency in the estimate of a total (total crop, total revenue, total sales, number of people). Schemes for automatic correction of the frame, such as use of the half-open interval in sampling a list, use of the solid segment of dwelling units with the half-open interval in demographic studies, give every unit a prescribed probability, even though the frame is out of date (as nearly every frame is), or was compiled carelessly. (Omission of whole areas, or of a whole tape, or of a whole file, is a different matter and can be corrected only by adding them to the frame.) Problems in the 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.

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, 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 only 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 that are fully mechanized and from platforms that are partially mechanized, and from platforms 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.

Most of the above points are obvious to anyone that has never studied statistical theory.

Sampling for a rare characteristic. A problem sometimes encountered is to find the probability that the proportion of errors of a certain type in a large number of accounts, or of defects in a lot, is below a prescribed level. A problem of similar nature exists in seeds: if the proportion of some foreign variety or of some undesired weed is more than a prescribed fraction, the seed would be impaired and down-graded. The problem resembles the chemist's determination that a trace of impurity exists (as of carbon monoxide in the air, or of selenium dioxide in a portion of manganese dioxide C. P.), but at a concentration so low that only a wild and worthless estimate of the amount of the impurity would be possible. The problem that I shall describe here is one of dormant routes amongst the 3,000,000 freight bills issued in a year by a carrier of motor freight. A dormant route is one that a carrier has abandoned to competition. A dormant route would reduce the value of the motor company to a willing purchaser.

Let N be the number of freight bills in the frame, and let N_1 be a number so prescribed that if the number of freight bills between two points be less than N_1, the route may be declared dormant. In this instance, N_1 was set equal to 60, which would allow a little more than 1 trip per week. This number was prescribed by the lawyers engaged in the sale of the company. Let n be the size of sample required to reach the probability Q of failing to pick up a single bill from a route that is not dormant. In this case, Q was set by the lawyers at 1/5. That is, they would be satisfied with a sample that would find 4 out of 5 of the routes in the frame over which there were only 60 trips in the year. Then

$$Q = (1 - N_1/N)^n$$

$$\ell n\, Q = n\, \ell n\, (1 - N_1/N)$$

$$\doteq n\, N_1/N$$

$$\frac{n}{N} \doteq -\frac{\ell n\, Q}{N_1}$$

$$= -\frac{\ell n\, 1 - \ell n\, 5}{60} = \frac{\ell n\, 5}{60}$$

$$= \frac{1.60944}{60} = \frac{1}{37} = \frac{27}{1000}$$

A convenient way to draw the sample would be to pull out 27 freight bills with random numbers from every consecutive 1000 bills in the files, thus reaping the benefit, whatever it be, of any natural stratification in the frame.

More effort put forth on the mathematics, as by retention of more terms in the series for $\ell n\,(1 - N_1/N)$, and use of hypergeometric terms in place of the binomial, yields no refinement worth a further line.

Interpreting standards. Industry is beset today trying to understand the meaning of standards imposed by the government on fireproofing of curtains, mattresses, sleepwear for children, even high chairs; standards for tires, drugs, foods, and a host of other consumer goods; standards of emission and pollution. The trouble is that a manufacturer can not tell from reading the standard whether a single defective item found on the shelf is grounds for recall of the entire lot, however it be scattered over the country, or whether he must demonstrate due care in manufacture.

A destructive test deprives the manufacturer of his right to a second test. Hence, if a test is destructive, the only meaning that a standard can have is in respect to the quality control in the manufacturer of items like the one tested.

Due care in manufacture can be defined operationally only by records of quality control. This is a statement of momentous importance, which in my observation has not reached lawyers that defend manufacturers against suits for defective items.

The work of Dodge and Romig in acceptance sampling gave to the world meaning to consumer's risk, producer's risk, lot-quality, and to protection afforded by the AOQL.

Standards of lot-quality for which no consumer's risk can be computed, nor an AOQL, is no standard at all. This is unfortunately a characteristic of some of the government standards, drawn up with good intentions, and in most cases with depth of insight in almost anything but the requisite statistical knowledge.

It is unfortunate that industry has failed to come through with appropriate standards for themselves, and are now at the mercy of standards issued by the government. Words of Senator Flanders (deceased), written in 1951, are prophetic.*

* Ralph E. Flanders, HOW BIG IS AN INCH? (Atlantic Monthly, Jan. 1951).

There are trends, plans, and proposals currently under way that would make standardization wholly or mainly a function of government, and I am opposed to them. I do not want my talented, capable, and sincere friends in the federal agencies in Washington to write the industrial standards of this country. Too much is at stake.

If you control an industry's standards, therefore, you control that industry lock, stock, and ledger. On the day that standards become a governmental function and responsibility, as is now being threatened, the government will take a very long step toward the control of American industry.

No government planner knows enough to write the standards for the rest of American industry and all other American people.

There is a group, both in and out of government, who feel that standardization is properly a government responsibility.

Last word. Some people would call this paper a view of sampling. Some would call it statistical theory in practice. Others would call it the statistical control of quality: others operational research: still others, systems analysis. To me, it is just some thoughts put forth by a statistician that is trying to be useful--pedestrian thoughts, but exciting and helpful in my own work.