Paper 142

SOME POINTS ABOUT STANDARD ERROR

By

W. Edwards Deming
Consultant in Statistical Surveys
4924 Butterworth Place
WASHINGTON 20016

An address delivered to the Market Research Council
in New York on the 17th September 1971

Introduction and purpose

This is the title that I gave to Mr. Dutka when, in a moment of weakness, I assented to talk to the Council today. He had asked everyone that he could think of, but they had all backed off with various excuses, so he turned to me in desperation.

Now everyone thinks that he knows all about standard error, and some people even have their own standards of error. I have little to offer to anyone that knows all about standard error. I may have something to offer to him that confesses that he could still learn a little, as I have been trying to, albeit the hard way, week after week, in 40 years of study and practice. Standard error is essential as a means of communication between research organization and client. We shall find that it is now easy to compute in studies of consumer research, and that the cost, thanks to computers, is small, but that it has meaning only in relation to proper statistical design. Its interpretation is more difficult, we shall see, than one would suppose by reading the books.

The purpose of this talk is to help people that do research, and people that pay for research, to get more for their money.

I am only a statistician, and I claim no knowledge about consumer-research, yet I do claim to appreciate its importance. Thus, in my own work, consumer-research is an indispensable ingredient in the control of quality of manufactured product. I may in this connexion refer you to lectures that I gave to Japanese management in 1950, when they first engaged me for work in the statistical control of quality, and in further work with Japanese industry on 11 subsequent visits to Japan. As I put it then and now, the consumer is the most important end of the production-line. What does quality mean except by watching the consumer's reactions? The whole world knows the results that Japanese management achieved through statistical methods applied at every stage of production, from procurement of materials to the consumer, with the aim of better design of product, with greater dependability, all at reduced cost. (Copies of the lectures on request.)

Much is new about statistical work. Improvement of response to sensitive questions by automatic randomization of a mixture of questions, for example, still in embryo stages, is fascinating. I wish that I could re-write now my own books in the light of new theory and new practice.

There are no secrets about statistical work, and no patents. Anyone can learn statistical theory if he has the determination to do so.

Does a figure convey information?

A figure may or may not convey information. It does only if we know what is wrong with it. If we don't know how far we can trust a figure, we should have a hard time to try to use it. We should remember John Tukey's remark that the more you know about what is wrong with a figure, the more useful it becomes. Thus, in modern practice, a competent research-organization, instead of claiming that every result that it produces is beyond question, serves its clients best by investing in appropriate statistical design and controls, to be able to evaluate the fringes of uncertainty in a figure before it leads to some costly business-decision.

What is the aim of good design?

I know nothing about management. I may say, however, from the viewpoint of a statistician, that it is difficult to understand how people can spend good money for research, or receive money for research, without having any idea of what they get or give for their money. Maybe the statistician is a better manager than most managers.

The aim of statistical design is to get the most information per unit cost (not the most figures, mind you), AND to provide information on which to evaluate the statistical reliability of the results. It seems to be known only to the statistician that it is possible to evaluate, by statistical methods, what allowance to make for sources of uncertainty described as Type II and III further on.

Be it noted that there can be no evaluation of the statistical reliability of a result unless the survey was properly designed, which means a probability-design and reasonable success in carrying out the field-work, including call-backs, plus careful coding, and calculation of estimates, all in reasonable conformance with the specifications.

Is statistical design good management? I think so. And the converse is so: good management requires statistical design of surveys.

Is good statistical work costly?

Some people have queer ways of counting money. The same man that spends $50,000 for research without any idea of what he gets for his money, presents a completely different behaviour when he buys a house. He has the title searched; he engages an expert to come and look the place over for termites, leaks in the roof, crumbling timbers, underground water, drains. If he builds a house, he hires an architect to be on the job to see that the specifications for excavation, footings, concrete, timbers, insulation, roof, painting inside and out, conform to the specifications as written.

The issue of cost has been clouded in the past by too many people that know so much that isn't so about statistical procedures, results, and costs, as I shall try to illustrate.

I realize that many of you would like to come back at me and say that you don't have money to spend on good research, or to furnish your clients with good research. The correct answer, to the man that is truly interested in costs, is that proper design and statistical controls reduce costs: they increase the amount of information per dollar, and they produce results that can be evaluated. I don't blame you for watching your pocketbook, but good statistical work is the best guardian thereof.

How can you know whether a result costs too much if you don't know the quality of the result? It is one thing to spend $10,000 for results of unknown quality, and another thing to spend $10,000 or $15,000 for results of demonstrable reliability. And by demonstrable reliability I do not necessarily mean high precision: I mean results whose uncertainties are evaluated.

I may point out, too, that in the Census, where the field-division knows where every dime goes, they have reduced costs, through statistical controls, including programs of reinterviewing, to the point where they pay about 1/3d as much per interview as is usual in commercial research, and they get quality incomparably better.

Can you distinguish a probability sample from something else?

Unfortunately, there is lots of research going on under the label of probability-sampling that is no such thing. It is important not to be fooled by words. I have heard of modified milk. No one knows how much water you can add, nor what else, nor how much cream you can take off, and still call it modified milk. I have heard of a modified probability-sample. This could mean anything, which means that the term conveys no information about what you get for your money. I have heard of procedures, called probability-samples, conducted by a quota-method in blocks that were selected by expensive probability-methods. I have also heard of a plan of starting the interviewer off at a designated point in such a sample of blocks, to continue until she has 10 interviews: then go to the next designated block.

These are not probability samples. They may be better than interviewing on the street corner, but how do you know? Such performance gives in advance no probability to anyone of being selected. It makes no record of vacancies, addresses not dwelling units, of people not at home, senile, with language barrier, nor other characteristics of nonrespondents. Such plans blissfully ignore nonresponse. They just sweep them under the rug. There is no record as a basis for improvement.

Why take a chance? The sad feature about such procedures is that in the selection of blocks it has already paid out 80% or more of the cost of preparation for a good job. The other 20% is used to ruin the foundation.

Another point is that the huge clusters or segments that some concerns use may be woefully inefficient even for products of low incidence. This would be obvious from a small amount of experimental work that could be woven into the regular surveys, were they conducted on a probability-basis, and were standard errors computed. The only way to arrive at the optimum size of segment and the correct allocation of a sample is through use of the appropriate statistical theory. The appropriate statistical theory involves variances between segments within counties and variances between means of counties, along with the costs of preparation, travel, and interviewing.

It is always possible to determine by questions and probes whether a procedure is a probability-sample, though the prescription of such tests may require considerable knowledge of theory and maturity in experience.

I give a number of lectures every year at the Graduate School of Business Administration on various topics in statistical methods. I am thinking of introducing a new title: HOW TO DOUBLE YOUR COSTS. However, such a lecture may not be necessary. Simple calculations that I make time to time indicate that some people are already pretty successful in doubling their costs without benefit of lectures.

Three types of uncertainty

In my own work, I have found the following classification of uncertainties in data to be helpful, as it indicates the sources of trouble.

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

1. Failure to perceive in advance what information would be useful.

2. Inept wording or sequences of questions.

3. Omission from the frame of important classes of the universe.

4. Unfortunate choice of date or 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.

8. Unwarranted inferences on the part of the user. Failure to take account of environmental conditions.

Type II. Operational blemishes and blunders

9. Small errors of a non-cancelling nature (e.g., omission of a sampling unit).

10. Persistent favor of an interviewer in one direction, causing an operational bias, and the like.

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

Type III. Random variation

12. Random variation arises from differences that exist by anybody's standards between households, segments, and other sampling units, and from other small accidental independent variations, such as the time of day for the interview, direction of travel, etc. The standard error is a universal measure of random variation.

The effect of structural limitations (Type I) can be evaluated by experimental design in which one changes the questionnaire or method of training or method of interviewing, or all three. The effect of the date chosen for a survey can be evaluated only by comparing the results with those obtained at another date, by the same methods. A re-canvass will not detect nor evaluate defects of Type I.

In contrast, uncertainties of Type II can be assessed by audits or statistical controls. Outside comparisons are sometimes helpful--that is, comparison with the Census or with other surveys (though comparison requires considerable maturity, as one almost never finds a survey sufficiently identical with the Census or with any other survey to permit useful comparisons).

Standard errors, which measure the uncertainties of Type III, can be evaluated provided the survey was laid out properly. In fact, statistical design of a survey specifies, as a matter of course, the procedures by which standard errors are to be computed. The standard error wraps up all the random variations, including the differences between interviewers. There is, in addition, in statistical design a separate evaluation of the variance between interviewers, at least in the large cities.

Variances between interviewers

Variances between interviewers (the combined effect of the question, the respondent, and the interviewer) can be evaluated at low cost in large cities, so that the questionnaire can be improved and the interviewing likewise. Without statistical tests, one remains blissfully ignorant of the contributions to bias and to variance that arise from the questions and from the interviewing.

I exhibited three years ago before this Council some results to show how important the interviewer is in some kinds of questions. Almost any manufacturer of an article of food or of laundry-equipment used in the home would wish to know how many female homemakers ever heard of his product. The sad truth is, though, that answers to the question, "Have you ever heard of Brand X of (e.g.) floor-wax?" are plagued with terrific variance, almost certainly from interaction. Nor can one rely on answers about the usage of products that generate embarrassment.

Figures show that for many questions of these types, the effective size of sample is about equal to the number of interviewers, which points to the strong possibility that the interviewers don't ask the question at all, but just put down their own ideas. For such questions, in the present state of the art of questioning, the sensible way to conduct the survey is to ask the interviewer to fill out the answers at home, and don't charge us for travel.

One of your clients remarked to me one day when I mentioned the possible magnitude of differences between interviewers, "Why not require all the interviewers to use the same form for the questions?" Unfortunately, the problem is not that simple. Everyone uses a form and requires the interviewers to follow it rigidly, yet (though) he may not know it, he has problems.

One might suppose that age is definite; that the simple question, "How old are you?" could have only one response, yet in an actual trial on this very point, 10% of the ages recorded for 8500 persons, on a re-interview only 7 days later, turned out to be different by a year or more. Out of 300,000 men who recorded their ages on 2 legal documents, 2 years apart, 9% differed by more than 2 years, and 2% differed by 5 years or more.

Duplicate reports on the occupation of 4500 workers, one report coming from the worker himself or some member of the household, and the other coming from the worker's employer, showed 22% discrepancy when the figures were tabulated by 9 broad occupation groups. The discrepancy rose to 35% when the data were tabulated in finer groups.

Nothing much will happen in the improvement of the questionnaire and the field-work so long as the problem is regarded as simple. Standardized or no, differences exist and they are huge. They will remain so until research organizations become serious enough to do something about them. With so much variance coming from the interaction between respondent, questionnaire, and interviewer, one would suppose that research organizations and their clients would conduct tests on questions to try to reduce this variance.

Your clients worry about trends. They should, if they are good businessmen. Let me point out, however, that an alarming jump up or down in two successive surveys could be caused by a turn-over of 25% of the interviewers, or by changes in supervisors or in changes in methods of supervision. Analysis of data requires measurement of possible effect from the interviewer. Analysis of variance, so widely tooted in books and in courses, is important indeed, but it is only one tool in the analysis of data. Nothing takes the place of a plain scatter-diagram, nor of a good look at the differences obtained by old interviewers and by new interviewers, even though this examination is plagued with the confounding of area and interviewer.

I would regard variance between interviewers, when it is large, as a structural defect, as the trouble lies in the questionnaire and in the head office of an organization that is satisfied with failure to evaluate supervision and training.

Use of the standard error

One of the main uses of standard errors is that it detects the need for improvement of questionnaires and field-work. Standard errors then measure the effect of changes--measure of improvement, if any.

Interpretation of the standard error in terms of the margins of uncertainty for a given probability such as odds 9 to 1 or 19 to 1 are not as simple as one would suppose by reading books. In the first place, one should examine the results of the sample for extreme skewness. Any small proportion

such as .05 will present problems of skewness, though large proportions may also exhibit skewness, signifying huge geographic differences, or a heavy impact from new interviewers. Simple transformations will often remove the skewness, but this is not the place for a mathematical treatment.

It is also important here to examine the meaning of a standard error when skewness is not a problem.

Oversimplified, the standard error of a result will consists of three terms. Thus, for an estimated proportion $\hat{p}$,

$$\sigma^2_{\hat{p}} = \frac{A}{m} + \frac{B}{n} + \frac{C}{k}$$

where m is the number of primary sampling units in the sample, n the number of people, k the number of interviewers, supposedly randomized.

What do we mean by a lower confidence limit for odds 9 to 1? It would depend on whether we had randomized the interviewers. There would be one interpretation if we had drawn by random numbers the interviewers used from a pool of eligible interviewers. This is usually not the case. More usually, the interviewer and the county are bound together. There is one good interviewer, and we hire her. Interpretation of the lower confidence limit of the 9 to 1 would then be that we greatly increased the size of the sample, retaining the same number of interviewers, increasing their work-loads proportionately, the result would not fall below the limit calculated. The odds of the statistician being wrong on a confidence interval properly calculated would be 9 to 1.

A standard error is a mark of quality, but a meaningful standard error can be calculated only if the sample was designed properly, and carried off in reasonable conformance with the specifications. This means designation of households or solid segments of households for interview, with a definite probability of selection specified in advance. It means call-backs and call-backs on people not previously at home. And when I say call-backs, I mean directed call-backs, with information in advance on when to call: not call-backs at random times, nor call-backs billed but never made. If you think that call-backs are too expensive, look at the figures in Fable 2.

Here in a nutshell are some points to remember about a standard error:

1. It is an honor for a result to have a standard error. It is a mark of quality.

2. Very few surveys attain to this honor.

3. Every result, if it was worth paying for, is worthy of a standard error.

4. A standard error has no useful meaning unless the sample-design is a probability-sample reasonably well carried off in the field. A standard error of a result has no use unless the result itself will be useful.

5. The procedure for computation of a standard error depends on the sample-design (see Fable 4).

6. Standard errors arise from all the small independent accidental variations, such as differences between households by anybody's standards, differences between interviewers, direction of travel, time of day for the interview, and myriads of other random contributions.

Of these contributions to the standard error, the contribution from the differences between interviewers, through fault in the construction of the questionnaire, and because of faults in the field-work, unfortunately for many questions outstrips all the other contributions.

For example, I demonstrated to the Market Research Council assembled here three years ago that questions such as, "Have you ever heard of _____," are practically useless, because of the huge variance between interviewers. You get the interviewer's idea on the answer: not the respondent's.

Unless these contributions to the variance are measured, one has little idea how to interpret a result, nor on how to improve the questionnaire and the field-work.

Details of sampling procedure are unimportant in a report to the user

Details concerning the sampling procedure of a probability sample are of little or no help to the user of the data. In statistical language, the sampling procedure contains no new information, once the user has in hand a careful evaluation of the statistical reliability of the results of a probability sample. Specifically, what constituted the sampling unit, the stratification used, the method of selection, the formula for the calculation of estimates, and the procedures for estimating standard errors, yield no new information about the reliability of a result.

It is customary in some circles to present results in a beautiful book with a technical appendix to tell about the sampling procedure. Such an appendix may be impressive, but if the procedure was a probability sample, it is no help to the user: it adds no new information to a thorough-going presentation of statistical reliability.

It is customary, however, and justifiable I believe, to present in a legal case a copy of the sampling procedures and instructions for coding, along with calculations, and the results of probes, so that anyone interested may investigate the actual work done to see if it conformed to the procedures. Opposing counsel has not only a right but a duty to satisfy himself that the actual performance met the specifications, or fell short.

Fables about surveys

Fable 1. Good statistical work is costly.

WRONG. Good statistical work means, by definition, maximum information per unit cost. One can not argue intelligently on the costs of surveys without (a) figures on costs, and (b) criteria on which to judge the quality of the results.

The quality of a result is assessed by its statistical reliability. (I did not say usefulness: a result may be highly reliable but useless.) Any shoddy job can be replaced by one that is shoddier and cheaper. It is no great achievement to cut costs. There is a simple way to save the whole cost of a proposed study: don't do it at all.

It is a fact that one can carry out a respectable probability sample, with call-backs, at less cost than some designs that I have seen that are deplorable, to put it mildly.

Fable 2. Call-backs are costly.

WRONG. This fable has its origin in tall talk without figures. I am reminded of the Book of Job, Chapter xxxvii, 2d verse: "Who is it that darkeneth counsel by words without knowledge?" All interviews cost money, even interviews in the lobby or on the street corner. Thus, if one merely picks up what interviews he can find in a block, without call-backs, he will find, if he keeps records, that around 75% of the doors knocked on will be classified as not at home, not a dwelling unit, away for the duration of the survey, vacant for sale or for rent, not an eligible member of the universe (for example, no female head of household when the survey deals with female home-makers), language-barrier, senile, and once in a while a refusal. That is, three-fourths of the doors knocked on produce no results. All this knocking on doors takes time and costs money.

On the other hand, at the 2d call, if it is made with intelligence, the yield may well increase 30% to 50% over the yield at the 1st call. The overall cost per interview completed at the 2d call will be less than the cost per interview completed at the 1st call. The 3d and 4th calls, if properly supervised, will be increasingly productive.

Fable 3. One has a probability sample if he selects blocks or other areas with known probability; stops there; fails to designate the respondents.

WRONG. I have already, in earlier paragraphs, disposed of this fable. As I said, this kind of selection may be better than interviewing in the lobby or on the street, but I know of no tests that would substantiate such hope.

Fable 4. There is a standard table for standard errors, depending on p, q, and n.

WRONG. There is no standard table of standard errors. Every question in a survey begets a result, and every result has its own separate computation of standard error.

Thus, the proportions for two results might turn out to be the same, but the standard errors could nevertheless be different. For example,

	Proportion	Standard error
Female homemakers buy Brand X with regularity	.25	.05
They have heard of Brand Y	.25	.10

Possible reasons for the difference in standard errors are divers. The reasons could be different reactions between respondents and interviewers on the two questions, differences between households within segments, differences between segments within counties, or from differences between counties. (I am using county in the sense of a county, a group of counties, or a portion of an SMSA.) It is a fairly simple matter, especially now with computers, to separate out these effects and to know where the contribution is coming from. If we know where the contribution to variance comes from, we can sometimes do something about it.

Fable 5. The more interviews in a survey, the more accurate the results. If we specify 1000 interviews in a survey, we know in advance the reliability of the results. A large number of interviews will cover up defects in procedure.

WRONG. The effect of structural limitations is the same for a big sample as for a small one. Size does not shrink defects. Moreover, the number of interviews is not even useful as a determinant of the standard error (Fable 4). What counts equally is the procedure of selection and the procedures for calculation of estimates. Moreover, the nonsampling errors may actually increase as we increase the size of a sample.

A thousand interviews mean nothing without specification of the entire procedure, including the methods of estimation that will be used to produce the results. Some organization can always conduct 1000 interviews cheaper than somebody else when there are no specifications and hence no test possible: anything will do. The man that merely specifies 1000 interviews without specifications is almost sure to get rooked.

Research should be bought and sold on the basis of statistical reliability.

Fable 6. A sampling expert is a man that selects a part of a frame from the whole.

WRONG. A sampling expert is a man that guards your pocket-book. He does this by using statistical theory through all stages of survey-work, including (a) tests of questions; (b) statistical controls to improve field-work and to detect and evaluate non-sampling errors (Type II): (c) analysis of data, to measure the effects of non-sampling errors, effects of the differences between the effects of non-sampling errors, effects of the differences between interviewers, as well as uncertainty from random variation.

The size of sample and procedures that he specifies for calculation of estimates and of standard errors, and for controls to detect and to measure the effect of nonsampling errors, will be whatever appear to be the most economical to achieve the statistical reliability desired. Under certain circumstances, the optimum sample will be 100% of the frame.