STRAIN AT A GNAT AND SWALLOW A CAMEL: OR, THE PROBLEM OF MEASURING SAMPLING AND NON-SAMPLING ERRORS

Tore Dalenius, University of Stockholm

> "I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge; but you have scarcely, in your thoughts, advanced to the stage of a science, whatever the matter may be."
>
> -- Lord Kelvin

1. Introduction

The utility of a survey (program) may be expressed in terms of a "utility vector" $U(\cdot)$ with elements:

i. Relevance
ii. Accuracy
iii. Timeliness
iv. Wealth of detail

etc. This paper will focus on the two first-mentioned elements.

It is clear that measures of the relevance and the accuracy are most useful to both producers and users of survey statistics.

Reliable statistics on survey practice are unavailable. It is hoped that the forthcoming ASA-NSF "survey of surveys" will fill this gap in our present knowledge. Lacking such statistics, I will have to base this paper on "trade talk". The opinion appears to be widely held among statisticians and users alike that much of today's survey practice is inadequate for several reasons, three of which are as follows:

a. The relevance is seldom measured. The point seems to be that the relevance is usually taken for granted rather than objectively assessed. It is, for example, not necessarily true that concepts which were adequate 30 years ago, when a given survey program was started, are still adequate.

b. Too often, no (satisfactory) effort is being made to measure the accuracy. Statisticians are typically content to measure the sampling error, while neglecting the non-sampling error. The following quotation, from Wallis (1971), is illuminating:

> "Although there was considerable variation, both for different statistics in the same agency and across agencies, the [Commission's] response to the survey showed disappointingly little knowledge of error structure. Sampling errors were estimated for most statistics based on probability samples, but there were, with only few exceptions, very few analyses of response and other nonsampling errors, even in cases in which, because of long recall or the use of incomplete records, they were likely to be substantial."

c. Measures of the sampling error are too often grossly inadequate. Thus it is not uncommon to use a formula for simple random sampling irrespective of the sampling design actually used. Mention should also be made of the practice of neglecting the fact that what is referred to as a measure of the sampling error may also to some extent reflect response variation.

The survey practice just described may be summarized in terms of "strain at a gnat and swallow a camel"; this characterization applies especially to the practice with respect to the accuracy: the sampling error plays the role of the gnat, sometimes malformed, while the non-sampling error plays the role of the camel, often of unknown size and always of unwieldy shape. There are some signs today that the situation just discussed is worsening: non-response rates have increased significantly in recent years and may become even higher.

If today's unsatisfactory survey practice is not to become tomorrow's malpractice, a radical change is called for. It is the modest purpose of this paper to review the prospects for change and to discuss in general terms an approach to increasing our knowledge about the error structure of surveys, which in my opinion is a sine qua non for that change.

I. BRINGING ABOUT A CHANGE

2. The Notions of "Relevance" and "Accuracy"

The terminology used in discussions of relevance and accuracy is - as shown in Deighton et al. (1977) - characterized by a considerable amount of "linguistic variability", which makes the exchange of ideas and results difficult, to say the least. A necessary though not sufficient condition for bringing about a change is the development of a standard terminology. It is beyond the scope of this paper to suggest standards. I will be satisfied with defining "relevance" and "accuracy" along the lines suggested in Hansen et al. (1964).

The starting point is provided by three basic concepts:

i. the ideal goal $\bar{Z}$
ii. the defined goal $\bar{X}$; and
iii. the outcome of the survey $\bar{y}$.

Using these concepts, it is now possible to define three differences ("errors"):

i. $D(R) = \bar{X} - \bar{Z}$ (reflecting relevance)
ii. $D(A) = \bar{y} - \bar{X}$ (reflecting accuracy)
iii. $TD = \bar{y} - \bar{Z} = D(R) + D(A)$ (reflecting total difference)

3. A Plan for Bringing About a Change

The change I envision is to make it a rule, not an exception, that relevance and accuracy are measured.

It is clearly no easy matter to bring about such a change; it may call for many years of hard work. While there may be several alternative courses which would accomplish the same end, I will expand upon one specific one here.

The plan takes as its starting point (my conception of) the mechanisms (sources of errors) which generate the differences D(R) and D(A). Some resulting contributions to these differences are of a random nature and may thus be modeled by means of random variables and measured in terms of variances. Other contributions are of a systematic nature; they must be measured in terms of biases. The plan calls for reducing the biases even at the possible expense of increased variances. This idea is, of course, not new; it has long been used, for example, at the U.S. Bureau of the Census (Hansen et al. (1967)). The rationale of the plan is that it is usually much easier to cope with random errors than with systematic errors.

In sections 4 and 5, I will discuss how this plan may be applied to the control and measurement of the relevance and the accuracy, respectively.

4. Control and Measurement of the Relevance

Whether the statistics to be produced are classified as "general-purpose" or "special-purpose", the design of a survey must clearly reflect some specific purposes. The statistician must take into account who the potential users are and what the problems are to the solution of which they expect the survey to contribute.

The design procedure can indeed be formalized in a way that should enhance the control and measurement of D(R). I will dwell upon one such formalization here.

4.1 Control of D(R)

Consider a group of potential users with related or similar problems. Associated with this group, there is a set of ideal goals:

$$\bar{Z}_1, \bar{Z}_2, \ldots, \bar{Z}_j, \ldots, \bar{Z}_k$$

Corresponding to these ideal goals, there is a set of feasible defined goals:

$$\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_i, \ldots, \bar{X}_h$$

where typically $h \leq k$.

For each pair $\bar{X}_i$, $\bar{Z}_j$, there is a difference:

$$D_{ij} = \bar{X}_i - \bar{Z}_j$$

which reflects the relevance of $\bar{X}_i$ vis-à-vis $\bar{Z}_j$. This difference may be exhibited as a matrix:

$$[D_{ij}] = \begin{bmatrix} D_{11} & \cdots & D_{1j} & \cdots & D_{1k} \\ \vdots & & \vdots & & \vdots \\ D_{i1} & \cdots & D_{ij} & \cdots & D_{ik} \\ \vdots & & \vdots & & \vdots \\ D_{h1} & \cdots & D_{hj} & \cdots & D_{hk} \end{bmatrix}$$

If this matrix were known prior to the survey, it could be used to select the defined goal which in some sense is best. But by the same token, there would then be no need for the survey!

What is needed, obviously, is some method for approximating the matrix.

4.2 Approximating $[D_{ij}]$

I will point to two possible ways of approximating $[D_{ij}]$. One way calls for replacing the elements D_{ij} by "preference scores" P_{ij} reflecting the preferences among the users as to the pairs $\bar{X}_i$, $\bar{Z}_j$. Another way calls for replacing the elements D_{ij} by indicators a_{ij}, where $a_{ij} = 1$ if the pair $\bar{X}_i$, $\bar{Z}_j$ is judged by the users to be acceptable, and otherwise $a_{ij} = 0$.

A matrix $[P_{ij}]$ or $[a_{ij}]$ can be analyzed and assessed as a basis for selecting the defined goal (which is not necessarily one of the originally conceived defined goals). The analysis and assessment may be carried out along the lines discussed in Dalenius (1968). A condition for this procedure to be feasible and useful is obviously that intimate cooperation be established between the statistician and the users. The statistician must take an active role in getting to understand the users' problems; by the same token, the users must learn to understand the ramifications of alternative choices of defined goal. In addition, the computing expert must take an active part in the design.

In some cases - perhaps more often than not - it may not be feasible to approximate $[D_{ij}]$ as suggested above. In these cases, the construction of an error profile, to be discussed in part II, may provide some helpful insights.

5. Control and Measurement of the Accuracy

The difference D(A) may be written:

D(A) = sampling error + non-sampling error

where the sampling error is relative to the outcome of "equal complete coverage" (Deming (1960)) and the non-sampling error accounts for the balance of D(A).

5.1 Control of D(A)

While control should be aimed at both components of D(A), it seems especially important to focus on the non-sampling component. The error profiles presented in Bailar and Brooks (1977) and Madow (1977) support that contention. In what follows, I will discuss two possible approaches to control of the non-sampling component.

a. The first approach calls for identifying survey operations with high risks for deviation between design and execution which are difficult to control in a satisfactory way; (some of) these operations may then be replaced by operations with low risks. In some cases, this will mean replacing a "complicated" operation by a "simple" one, especially if the complicated operation is primarily a human operation (like coding). In other cases, the action to take will be the reverse one: a "simple" human operation is replaced by a "complicated" automatic (computerized) operation. Editing is an example of an area in which this idea is already successfully applied.

b. The second approach calls for (better) monitoring of the survey operations. This may necessitate the development of a special signal system which helps to identify problems while there is still time to take "preventive action". Non-response is an example of a kind of problem in which this approach should be relatively easy to apply in all surveys.

5.2 Measurement of D(A)

It is worth noting that theory and methods are in fact available for this measurement.

a. In the context of what has become known as the U.S. Bureau of the Census survey model, theory is developed for measuring D(A) by the mean-square error of the estimator:

MSE = sampling variance + response variance + interaction + squared bias

as discussed in Hansen et al. (1964). Moreover, methods are available - see Bailar and Dalenius (1969) for a systematic account - for the design of schemes which may be used to estiamte the components of the MSE.

b. In Lessler (1974), theory and methods are available for using two-phase sampling as a means of controlling and measuring D(A).

Thus, the fact that D(A) is seldom (adequately) measured cannot be explained by lack of the tools necessary for doing so. While there may be many reasons for the current state of affairs, I presume that cost considerations often play a decisive role.

At any rate, it is likely that a change will not take place spontaneously; it will have to be generated. One way of stimulating the change may be to illuminate the importance of measuring D(A) by means of some "second-best" measurements:

i. Measuring the "representativity" by comparing survey estimates with known population characteristics. This type of measurement dates back to the early days of purposive selection, and was often used in an uncritical way.
ii. Applying "error ratio analysis" as suggested by Brown (1967).
iii. Computing "quality codes" as developed by Zarkovich (1967).
iv. Constructing an "error profile".

In part II, I will dwell on this fourth option; I will in fact argue that it may serve a useful purpose with respect to both D(R) and D(A).

II. THE ERROR PROFILE APPROACH

6. The Notion of an Error Profile

Hansen et al. (1967) discuss what to do in a situation in which it is not feasible to measure D(A) by means of the mean-square error. In essence, they suggested that the statistician provide for a disclosure of the survey operations.

The term "error profile" will be used here in a way consistent with that suggestion; this term is chosen in preference to the longer though somewhat more appropriate term "profile of sources of errors". More specifically, an error profile is a systematic and comprehensive account of the survey operations which yield the statistic $\bar{y}$ and thus the differences D(R) and D(A).

Constructing an error profile calls for assessing each survey operation with respect to:

i. The presence or absence of a deviation between design and execution.
ii. The size of this deviation.
iii. The impact of this deviation; as a special case, this impact may be expressed in terms of a contribution to the MSE.

It should be remarked that it may not be possible to assess each survey operation with respect to all three elements just listed.

There is no standard format for an error profile. I will mention here two possible formats:

i. One format is based on a list of the survey operations in the order in which they were executed.
ii. The second format calls for assigning the survey operations to homogeneous groups on the basis of the purpose of each operation.

The second format has, it seems to me, the advantage of lending itself to some standardization. In section 7, I will present one possible grouping scheme.

7. A Possible Grouping of the Survey Operations

The starting point is the total difference:

$\bar{y} - \bar{Z} = D(R) + D(A)$

Against the background of this difference, I identify a hierarchy of survey operations:

i. Primary Survey Operations - PSOs
ii. Secondary Survey Operations - SSOs
iii. Tertiary Survey Operations - TSOs, etc.

In the interests of being specific, I will give a couple of illustrations.

It seems natural to distinguish two PSOs:

PSO-1: Design of the survey
PSO-2: Execution of the design

PSO-1 may be divided into SSOs as follows:

SSO-11: Choice of the properties to measure
SSO-12: Choice of the survey population

while PSO-2 may be divided into SSOs as follows:

SSO-21: Getting observational access to the population: developing the frame, selecting the sample, etc.
SSO-22: Collecting the data
SSO-23: Processing the data
SSO-24: Computing the survey statistic $\bar{y}$
SSO-25: Computing measures of D(R) and D(A)

8. The Assessment Procedure

As mentioned in section 6, constructing an error profile calls for assessing each survey operation with respect to three aspects: presence/absence of a deviation between design and execution; size; and impact. The procedure to use for this assessment will, of course, depend upon the nature of the survey operation to be assessed. I will limit myself here to giving two minor illustrations.

8.1 Assessing PSO-1: Design of the Survey

As discussed in section 7, PSO-1 may be divided into two SSOs:

SSO-11: Choice of the properties to measure
SSO-12: Choice of the survey population

I will discuss the assessment of these SSOs in turn.

a. Corresponding to the defined goal $\bar{X}$, there is a survey variable X defined by reference to:

i. the property to measure
ii. the measurement method

Similarly, corresponding to the ideal goal $\bar{Z}$, there is an ideal variable Z defined in the same way.

The analysis of the choice of the survey variable calls for determining whether the survey variable is equal to the ideal variable with respect to "property to measure" and "measurement method": X = Z, or whether it differs from the ideal variable: $X \neq Z$.

In a specific survey, X and Z may be defined by reference to the same property, but they may differ with respect to the measurement method. As an example, the measurement method corresponding to Z may not be operationally feasible for the survey under consideration.

If the analysis shows that $X \neq Z$, there is a "definitional bias" associated with the defined goal $\bar{X}$.

b. Corresponding to the defined goal $\bar{X}$, there is a population of objects - the survey population - to be denoted by $[O(\bar{X})]$; technically, it is represented by the frame.

Similarly, corresponding to the ideal goal $\bar{Z}$, there is a population of objects - the target population - to be denoted by $[O(\bar{Z})]$.

If all objects in $[O(\bar{X})]$ are also in $[O(\bar{Z})]$, and all objects in $[O(Z)]$ are also in $[O(X)]$, the survey population is equal to the target population:

$$[O(\bar{X})] = [O(\bar{Z})]$$

If some objects in $[O(\bar{X})]$ are not in $[O(\bar{Z})]$, or some objects in $[O(\bar{Z})]$ are not in $[O(\bar{X})]$, the survey population is different from the target population:

$$[O(\bar{X})] \neq [O(\bar{Z})]$$

In fact, this latter situation is the typical one in applications. It calls for assessing the difference between $[O(\bar{X})]$ and $[O(\bar{Z})]$ by comparing these populations with respect to:

i. The rules associating objects with $[O(\bar{X})]$ and $[O(\bar{Z})]$, respectively;
ii. The (approximate) frequencies:

N_{11} = the number of objects which belong to both $[O(\bar{X})]$ and $[O(\bar{Z})]$
N_{10} = the number of objects which belong to $[O(\bar{X})]$ but not to $[O(\bar{Z})]$
N_{01} = the number of objects which belong to $[O(\bar{Z})]$ but not to $[O(\bar{X})]$

The ratio:

$$R = \frac{N_{11}}{N_{11} + N_{01}}$$

may be looked upon as a measure of the appropriateness of $[O(\bar{X})]$.

The point just made about R may be illustrated by considering the case of a survey which yields an estimate $t = pN_{11}$ of the target population total:

$$T = P(N_{11} + N_{01})$$

where P is, for example, the rate of unemployed persons. If R is close to 1, then t is close to T (granted that p is close to P).

8.2. Assessing TSO-221: Observing the Objects Selected for the Survey

In real-life surveys, the number of TSOs is likely to be large. I will select one of them - TSO-221 - for illustration: observing the objects selected for the survey (irrespective of the method of operation).

In practice, it will happen that some objects become "non-respondents". It is in principle relatively simple to measure the size of this specific event; this does not mean, however, that it is adequately done in all instances. As to measuring the impact of the non-response, it is in some cases (notably when $\bar{X}$ is a proportion) possible to compute an upper and lower value for this impact.

9. The Error Profile Documentation

A comprehensive account of the assessment of survey operations may possibly become a rather sizeable document, especially if it is to be self-contained and deals with a survey which is not repeated. It may therefore prove desirable to try to summarize these findings in a simple error profile protocol, or table, the headings of which may be as in figure 1 below.

Survey operation	Kind of deviation	Size	Impact
PSO-1: Design of the survey			
SSO-11: Choice of properties to measure			
SSO-12: Choice of survey population			
PSO-2: Execution of design			
PSO-21: ...			
PSO-22: ...			
etc.			

Figure 1.

10. Limitations and Potentialities of the Error Profile Approach

In sections 4 and 5, constructing an error profile was suggested as a means of measuring D(R) and D(A).

In section 6, I defined an error profile to be "a systematic and comprehensive account of the survey operations which yielded the statistic $\bar{y}$ and thus the differences D(R) and D(A)."

The error profile approach is as yet virtually untested. Thus, it would be premature to pass

any judgment on its usefulness; the proof of the pudding is in the eating.

The main limitation of the error profile approach is obvious: it does not make it possible to measure the components of the mean-square error of $\bar{y}$ within the framework of some survey model. On the other hand, the limited experiences as yet available support the contention that it has some significant potentialities. Thus, the error profile approach:

i. encourages comprehensive documentation of the survey operations;
ii. helps to identify "error-prone" survey operations; and
iii. serves as a summary protocol of research and development already carried out and yet to be carried out.

Acknowledgement: This paper reflects in several ways discussions within the Subcommittee on Non-Sampling Errors, organized by the Federal Committee on Statistical Methodology, Statistical Policy Division (Office of Management and Budget).

References

Bailar, B.A. and Dalenius, T. (1969): Estimating the response variance components of the U.S. Bureau of the Census' survey model. Sankhya, B, 341-360.

Brooks, C.A. and Bailar, B.A. (1977): An error profile: employment as measured by the current population survey. Paper presented at the 137th Annual Meeting of the American Statistical Association, Chicago, Ill., August 15-18, 1977.

Brown, R.V. (1967): Evaluation of the total survey error by error ration analysis. Metra, 593-613.

Dalenius, T. (1968): A feasible approach to general purpose sampling. Management Science, 110-113.

Deighton, R.E., Poland, J.R., Stubbs, J.R. and Tortora, R.D. (1977): Glossary of nonsampling error terms. Paper presented at the 137th Annual Meeting of the American Statistical Association, Chicago, Ill., August 15-18, 1977.

Deming, W.E. (1960): Sample design in business research. Ch. 4. John Wiley & Sons, Inc., New York.

Hansen, M.H., Hurwitz, W.N. and Pritzker, L. (1964): The estimation and interpretation of gross differences and the simple response variance. In Rao, C.R. (editor): Contributions to statistics presented to Professor P.C. Mahalanobis on the occasion of his 70th birthday. Pergamon Press, Oxford, and Statistical Publishing Society, Calcutta.

Hansen, M.H., Hurwitz, W.N. and Pritzker, L. (1967): Standardization of procedures for the evaluation of data: measurement errors and statistical standards in the Bureau of the Census. Paper presented at the 36th Session of the International Statistical Institute in Sydney, 1967.

Lessler, J.T. (1974): A double sampling scheme model for eliminating measurement process bias and estimating measurement errors in surveys. Institute of Statistics Mimeo Series No. 949, University of North Carolina, Chapel Hill.

Madow, L.H. (1977): An error profile: employment as measured by the current employment statistics program. Paper presented at the 137th Annual Meeting of the American Statistical Association, Chicago, Ill., August 15-18, 1977.

Wallis, A. (Chairman) (1971): The President's Commission on Federal Statistics. Vol. II. Government Printing Office, Washington, D.C.

Zarkovich, S.S. (1967): A system of statistical quality codes. Paper presented at the 36th Session of the International Statistical Institute in Sydney, 1967.