MEASUREMENT ERRORS IN CENSUSES AND SURVEYS

By: Morris H. Hansen, William N. Hurwitz, and Max A. Bershad
Bureau of the Census

SUMMARY

Introduction

In a census or a sample survey we may obtain observations through personal inquiry, direct questionnaire, or other methods. The set of measurements or observations recorded in the collection operation ordinarily are examined for internal consistency and acceptability, certain "corrections" may be made, and some of the entries may be coded to identify them in a classification system. The results are then summarized into totals, averages, correlations, or other statistical measures. Taken together the collection and processing operations constitute the measurement process and are the source of any measurement errors.

In considering measurement errors we shall regard a survey as being conceptually repeatable, such that repetitions may relate to the same point or period in time and such that carrying through the operation once does not influence results obtained in repetitions of the operation. A particular survey result or estimate is the result of one trial. This conception provides the basis for defining a variance and bias due to response, processing, or other sources of measurement errors. Such a postulation can reasonably approximate actual conditions for a single survey, even though in practice independent repetitions of the survey may be impracticable or impossible.

Measurement errors may arise from many different types of causes, and depend on the conditions under which the survey is taken. Some of the conditions under which a survey is taken may be beyond the control or specification of the survey designer. Other conditions can be controlled so as to influence the quality of survey results in the sense that various aspects of the conduct of the survey are specified. These are typically imposed or attempted to be imposed in an effort to insure adequate quality. Conditions subject to control in survey design but which might be regarded as varying between the conceived repetitions of the survey are the particular choices of interviewers (if it is an interview survey) and other personnel chosen to do various aspects of the work, the particular assignments each is given, and other similar variable factors. These conditions in the conceived repetitions of the survey determine the averages, variances, correlations, and other expected values of various functions of the individual measurements.

The survey may be either a complete census or a sample, and a particular survey is regarded as one trial. For simplicity we shall consider the case of estimating a proportion from the survey. An observation on the j-th unit in the survey is designated by x_{jt}, which has the value of 1 if the j-th unit is assigned to the particular class under consideration on the t-th trial, and otherwise has the value 0. An estimate of the proportion of the population having a characteristic from a survey or trial is

$$p_t = \frac{1}{n} \sum_{j}^{n} x_{jt} \qquad (1)$$

where n is the number of units in the sample (with n = N for a complete census).

We have, then, the total mean square error of the estimated proportion is

$$MSE_{p_t} = E(p_t - p_s)^2 + E(p_s - P)^2 + 2E(p_t - p_s)(p_s - P) + B^2 \qquad (2)$$

where

$$p_s = \frac{1}{n} \sum_{j}^{n} P_j \qquad (3)$$

is the mean, for the particular set of units included in the s-th sample, of the P_j, where P_j represents the expected value of the observed results for the j-th unit over all possible results and samples in which the j-th unit appears, and

$$P = \frac{1}{N} \sum_{j}^{N} P_j$$

is the average result over all units and all possible observations under the conditions under which the census or survey is taken.

The first term in Eq. (2) is defined as the response variance contribution to the total variance of p_t, i.e., the response variance of p_t is defined as

$$\sigma^2_{\bar{d}_t} = E(p_t - p_s)^2 = E\left(\frac{1}{n} \sum_{j}^{n} d_{jt}\right)^2 = E\,\bar{d}_t^2 \qquad (4)$$

and is a function only of the response deviations, $d_{jt} = x_{jt} - P_j$. In the case of a complete census $p_s = P$, the second and third terms are zero, and this response variance term is the total variance of a census average.

The second term in Eq. (2) is defined as the sampling variance of p_t. When the indicated expected value of the second term is taken this becomes the usual sampling variance formula for the appropriate sample design as given else-

where.[1] The P_j's are the unique values associated with the units, j=1, ..., N, assumed in the usual theory for sampling from finite populations. In the case of a complete census, $p_s = P$ and the sampling variance term becomes zero.

The third term in Eq. (2) is twice the covariance between $\bar{d}_t$ and p_s. This term is not necessarily equal to 0. It will be zero for a complete census, or when repetitions of a survey are defined only for a fixed sample of units, and we shall ignore the effect of this covariance term in this paper.

The final term is the square of the bias of the survey estimate.

Correlated and Uncorrelated Response Deviations

The response variance given in Eq. (4) can be restated in the following form which separates the effect of uncorrelated and correlated response deviations:

$$\sigma^2_{\bar{d}_t} = \frac{1}{n}\sigma_d^2 [1 + \rho(n-1)] \qquad (5)$$

where

$$\sigma_d^2 = E\, d_{jt}^2 = \frac{1}{N}\sum_j^N P_j(1 - P_j) \qquad (6)$$

is the variance of the individual response deviations over all possible trials, and

$$\rho = \frac{E\, d_{jt}\, d_{kt}}{\sigma_d^2} \text{ (for j not equal to k)} \qquad (7)$$

is the intraclass correlation among the response deviations in a survey or trial.

The response variance contribution from uncorrelated response deviations is less, often in practice much less, than $(PQ)/n$, and if there are important contributions to response variance, they arise from the factors involving correlated response deviations.

The possible impact of even a very small intraclass correlation is substantial, as can be seen from an examination of Eq. (5). For example, if the intraclass correlation among response deviations is zero, the response variance of p_t is σ_d^2/n. Suppose, on the other hand, that the intraclass correlation is, say, .01 (a correlation so small that it might in other applications be regarded as of no consequence whatever). Suppose, also, that the sample or census involves the enumeration of n = 3,000 cases. In this case, the impact of the correlation is to increase the response variance by a factor of $\rho(n-1) = .01(2,999) = 30$, or 3000 percent! Thus, even if the response variance with uncorrelated response deviations is relatively small, when multiplied by such a factor it may be quite large.

As an example, an interviewer's misunderstanding of his instructions, carelessness, or a tendency to introduce his own judgments into a survey, may cause his results to differ from those of other interviewers, and thus be a source of correlated response deviations. A supervisor's interpretations of instructions that are passed on to interviewers under his jurisdiction and that differ from those of another supervisor may be another cause of correlated response deviations; the varying interpretations of different coders or other processors may be another cause.

We have carried through experimental studies that provide approximate rough estimates of the various correlated and uncorrelated response variance contributions to the MSE of estimates of various items in the 1950 Census. Our estimates can, in fact, be shown to be such that they will lead to understatements in general, of the various terms of the MSE, but are rough approximations and have proved useful in guiding census plans and further research.

As an illustration, we shall choose the proportion of persons classified as farmers and farm managers, which for April 1950, was estimated to be .039 in the 1950 Census. The corresponding estimated proportion from the Current Population Survey was .042. The difference of .003 is our estimate of the bias in the Census for this item (a relative bias of about 7 percent).

We shall now develop an approximation to the total mean square error for this illustrative item by making certain additional assumptions as to the repeated trials. We shall assume, first, that interviewers are independently selected and assigned in each repeated trial, but that other aspects of the staffing, procedures, etc., remain fixed. Also, we shall ignore any contributions to the response variance of correlated response deviations other than within the work of interviewers, and shall, as a consequence, understate the total response variance.

For this special case the total MSE can be written approximately

$$MSE_{p_t} \doteq \frac{\sigma_d^2}{n}[1 + \rho(\bar{\bar{n}} - 1)] + \frac{N - n}{N - 1}\frac{\sigma_S^2}{n} + B^2 \qquad (8)$$

where N is the total persons of the area, n is the number of persons in the sample (n = N for a complete census), and $\bar{\bar{n}}$ is the number of persons covered by each interviewer. Also, ρ is the intraclass correlation among response deviations within the work of an interviewer, σ_d^2 is the response variance per unit given by Eq. (6),

$$\sigma_S^2 = \frac{1}{N}\sum_j^N (P_j - P)^2$$

is the sampling variance per unit, and B is the response bias.

N and n depend on the size of the population of the area under consideration, and approximate values for other parameters relating to the illustrative item we selected above, i.e., proportion of persons classified as farmers and farm managers, are given below:

$\bar{\bar{n}}$ = 1000 for a complete census and = 250 for a 25 percent sample

$P = .04$ and $PQ = .0384$

$\sigma_d^2 = .13PQ = .005$

$\rho = .03$

$\sigma_S^2 = PQ - \sigma_d^2 = .033$

$B = .003$

The values for ρ, σ_d^2, and B are based on material presented in certain other sources. [2], [3], [4], and [5]

We shall assume that the general conditions of the survey are substantially the same whether the coverage is based on a 100 percent or on a 25 percent sample -- i.e., that the above approximate values will hold in either case.

The accompanying table compares results of a 100 percent and of a 25 percent sample survey, for populations of different sizes. The above values are assumed to hold in each case, and the numbers in the table were obtained by substituting the values assumed above in Eq. (8).

The following inferences are drawn from such results for items as that illustrated (this item was selected for illustration because it was roughly typical of a good many types of measurements in the census):

1. The combined sampling variance and the response variance contribute significantly to the MSE for very small tabulation cells.
2. The $\sqrt{MSE}$ with a 25 percent sample is not substantially greater than $\sqrt{MSE}$ for a complete census, even for the smaller cells.
3. The response bias is the important contributor to the errors of census statistics, especially for large tabulation cells.

Inferences such as these were important factors in the introduction of sampling and in the development of other modifications in census methods that have been introduced into the 1960 Population and Housing Censuses. However, much more research is needed, and extensive work is planned, to evaluate the effectiveness of the present and alternative methods.

We expect to report more fully on the appropriate theory and empirical results in a forthcoming paper. Extensive experimental work is being planned in the 1960 Censuses that will provide much fuller information than now available on response variances and biases, and the experiments to produce these results are described briefly in another paper in these Proceedings.

Size of population N	Complete Census						25 Percent Sample					
	n	$\sigma_{\bar{d}}^2$ ($\bar{\bar{n}}$=1,000)	σ_p^2	B^2	MSE	$\frac{\sqrt{MSE}}{P}$ (%)	n	$\sigma_{\bar{d}}^2$ ($\bar{\bar{n}}$ = 250)	σ_p^2	B^2	MSE	$\frac{\sqrt{MSE}}{P}$ (%)
1,000.....	1,000	.000 155	--	.000 009	.000 164	32	250	.000 169	.000 099	.000 009	.000 277	42
2,500.....	2,500	.000 062	--	.000 009	.000 071	21	625	.000 068	.000 040	.000 009	.000 117	27
5,000.....	5,000	.000 031	--	.000 009	.000 040	16	1,250	.000 034	.000 020	.000 009	.000 063	20
10,000.....	10,000	.000 016	--	.000 009	.000 025	13	2,500	.000 017	.000 010	.000 009	.000 036	15
25,000.....	25,000	.000 006	--	.000 009	.000 015	10	6,250	.000 007	.000 004	.000 009	.000 020	11
100,000.....	100,000	.000 002	--	.000 009	.000 011	8	25,000	.000 002	.000 001	.000 009	.000 012	9

REFERENCES

[1] M. H. Hansen, W. N. Hurwitz and W. G. Madow, Sample Survey Methods and Theory, John Wiley & Sons, Inc., New York, 1953.

[2] U. S. Bureau of the Census, The Post-enumeration Survey: 1950, (in press).

[3] A. R. Eckler and W. N. Hurwitz, "Response Variance and Biases in Censuses and Surveys," Bull. de L'Institut International de Statistique, Tome 36-2e Livraison, Stockholm, 1958, pp. 12-35. Paper presented at the 30th Session of the I. S. I.

[4] U. S. Bureau of the Census, The Accuracy of Certain Census Statistics with and without Sampling, Bureau of the Census Technical Papers -- No. 2 (in press).

[5] W. N. Hurwitz and M. A. Bershad, Self-enumeration with Follow-ups, Bureau of the Census (unpublished memorandum), April 1958.