MEASUREMENT OF RESPONSE ERRORS IN CENSUSES AND SAMPLE SURVEYS

G.J. Brackstone, J.F. Gosselin, B.E. Garton
Census Survey Methods Division

Madow [1968] has proposed a two-phase sampling scheme under which response bias can be eliminated from sample surveys by obtaining 'true' values for a subsample of the original sample. Often in cases of Censuses or ongoing surveys, the subsample data are not used to correct the main survey estimates but to assess their reliability. The main purpose of this paper is to present methods by which reliability estimates can be obtained when true values can be determined for a subsample of units.

1. INTRODUCTION

A sample scheme was proposed by Madow [1965] under which the response bias could be eliminated from sample survey estimates by obtaining 'true observations' for a subsample of the original sample. This is achieved by using the estimate of bias from the subsample to correct the original estimate.

There are some instances, however, where the subsample data are obtained for evaluation purposes after the survey data has been published. In this case the objective is the measurement of the overall reliability of previously published survey data for the purpose of

a) informing the user of the data of its reliability (allowing him to make adjustments to it if he wishes), and,

b) informing the survey-taker of the sources of error in the survey so that improvements can be made in future surveys.

The purpose of this paper is therefore to present estimators of the reliability of Census and sample survey data when true values can be obtained for a sample or sub-sample of cases.

Two approaches to this problem are considered. In the first case, we consider the problem under the usual 'response variance - response bias' framework. This represents the main part of the paper. The second approach ignores the above model and attempts to measure only the net error in the particular survey observed.

The results in this paper apply to any sampling scheme for the original survey and to any sub-sampling scheme for true values, provided only that estimators of sampling variance are available for the sampling schemes used.

2. APPROACH I

2.1 Response Error Model

The response error model is based on the concept of independent repetition. We will first treat the Census case (i.e. 100% enumeration). Suppose, hypothetically, that many independent 'trials' of the Census could be made under the same general conditions. The Census 'estimate' for a given category of interest would then follow a certain frequency distribution. A particular Census figure may then be considered to be a random observation from a distribution of all possible Census estimates.

Let $\bar{X}(t)$ denote the Census estimate obtained at trial t and let $\bar{\mu}$ denote the corresponding true mean. The usual statistical parameter used to measure the reliability of an estimate in this situation is the Mean Square Error (MSE) of $X(\bar{t})$:

$$\text{MSE}\ (\bar{X}(t)) = \underset{t}{E}\ (\bar{X}(t) - \bar{\mu})^2$$

$$= \underset{t}{V}\ [\bar{X}(t)] + B^2$$

where B denotes the bias of $\bar{X}(t)$,

$$\text{i.e.} \quad B = \underset{t}{E} [\bar{X}(t)] - \bar{\mu}$$

$$= \bar{X} - \bar{\mu} \quad \text{where} \quad \bar{X} = \underset{t}{E} [\bar{X}(t)]$$

The response variance of $\bar{X}(t)$ is defined as

$$\underset{t}{V} (\bar{X}(t)) = \underset{t}{E} (\bar{X}(t) - \bar{X})^2$$

In general, therefore, bias results from errors that tend to occur in one direction rather than another. For example, if an error was present in the instructions accompanying the Census questionnaire, errors would tend to be systematic in one direction. Hence, bias measures the average net effect of all these possible factors.

On the other hand, response variance measures the random component of the error. For instance, in the case of an ambiguous question, a self-enumerated person may give different responses on different independent trials. These types of error depend on unknown factors that are impossible to control and may vary from trial to trial (e.g. the frame of mind of the respondent, the fatigue of the interviewer).

The above discussion applies to any characteristic obtained from a Census. However, it can easily be extended to cover characteristics obtained from a sample survey. In this case, let $\bar{x}_s(t)$ denote the estimate obtained from sample, s, at trial t. The MSE $[\bar{x}_s(t)]$ can then be expressed as

$$\text{MSE} [\bar{x}_s(t)] = E [\bar{x}_s(t) - \bar{\mu}]^2$$

where the expectation is taken over all possible trials and samples.

Letting $\bar{X} = E\ (\bar{x}_s(t)) = \underset{s}{E}\ \underset{t}{E}\ (\bar{x}_s(t)|s)$

and $B = \bar{X} - \bar{\mu}$.

$$MSE\ [\bar{x}_s(t)] = E\ (\bar{x}_s(t) - \bar{X})^2 + B^2$$

$$= \underset{s}{E}\ \underset{t}{V}\ (\bar{x}_s(t)|s) + \underset{s}{V}\ \underset{t}{E}\ (\bar{x}_s(t)|s) + B^2$$

where the first term measures response variance, the second measures sampling variance, and the third measures bias.

2.2 Estimation

In the response error model described above three statistical parameters are defined, the mean square error, the response variance and the bias. Ideally we would like to obtain estimates of all three of these parameters in order to assess the level of both random and systematic errors and to obtain an overall measure of reliability.

Under the assumption that the true value of a sample characteristic is known for a random sub-sample of the survey sample, unbiased estimates of the MSE and bias are derived below. An unbiased estimator of the true proportion, $\bar{\mu}$, is also given following Madow [1965]. However, under this framework, it is not possible to estimate the total response variance.

Suppose true values are known for a random sub-sample, s', from s. Let $\bar{x}_{s'}(t)$ denote the unbiased estimate made from the sub-sample s' using the values observed on trial t, and let $\bar{\mu}_{s'}$ denote the corresponding estimate using the true values.

So $$\underset{s'}{E}\ (\bar{x}_{s'}(t)|s,t) = \bar{x}_s(t)$$

$$\underset{s}{E}\ \underset{s'}{E}\ (\bar{\mu}_s) = \bar{\mu}$$

Thus $\hat{B} = (\bar{x}_{s'}(t) - \bar{\mu}_{s'})$ is an unbiased estimator of B.

Consider first the estimator $(\bar{x}_{s'}(t) - \bar{\mu}_{s'})^2$. Letting E denote expectations over s', s and t (i.e. $E = \underset{t}{E}\ \underset{s}{E}\ \underset{s'}{E}$).

$$E(\bar{x}_{s'}(t) - \bar{\mu}_{s'})^2 = V(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) + [E(\bar{x}_{s'}(t) - \bar{\mu}_{s'})]^2$$

$$= V\{\underset{t}{E}\ \underset{ss'}{E}\ (\bar{x}_{s'}(t) - \bar{\mu}_{s'})|t\} + \underset{t}{E}\ \underset{ss'}{V}\{(\bar{x}_{s'}(t) - \bar{\mu}_{s'})|t\} + B^2$$

Secondly, let $v_s(\bar{x}_s(t))$ be a variance estimator such that

$$\underset{s}{E}\ [v_s(\bar{x}_s(t))|t] = \underset{s}{V}\ [\bar{x}_s(t)|t]$$

i.e. an unbiased estimator of the sampling variance over the survey sampling scheme for a given trial t.

Thirdly, let $v_{ss'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'})$ be a variance estimator such

$$\underset{ss'}{E}\ [v_{ss'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'})|t] = \underset{ss'}{V}\ [(\bar{x}_{s'}(t) - \bar{\mu}_{s'})|t]$$

Now, if we define $\hat{MSE}\ (\bar{x}_s(t))$ by

$$\hat{MSE}(\bar{x}_s(t)) = (\bar{x}_{s'}(t) - \bar{\mu}_{s'})^2 + v_s(\bar{x}_s(t)) - v_{ss'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) \tag{2.1}$$

then

$$E(\hat{MSE}\ (\bar{x}_s(t))) = \underset{t}{V}\ \underset{ss'}{E}\ \{(x_{s'}(t) - \bar{\mu}_{s'})|t\} + \underset{t}{E}\ \underset{ss'}{V}\ \{(\bar{x}_{s'}(t) - \bar{\mu}_{s'})|t\}$$

$$+ B^2 + \underset{t}{E}\ \underset{s}{V}(\bar{x}_s(t)|t) - \underset{t}{E}\ \underset{ss'}{V}\ \{(\bar{x}_{s'}(t) - \bar{\mu}_{s'})|t\}$$

$$= V_t E_s (\bar{x}_s(t)|t) + E_t V_s (\bar{x}_s(t)|t) + B^2$$

$$= V_{t,s}(\bar{x}_s(t)) + B^2$$

$$= MSE\ (\bar{x}_s(t))$$

Thus $\hat{MSE}\ (\bar{x}_s(t))$ given by (2.1) is an unbiased estimator of $MSE\ (\bar{x}_s(t))$.

In the case where the original survey is a Census, the middle term in $\hat{MSE}\ (\bar{x}_s(t))$ disappears and we have

$$\hat{MSE}(\bar{X}(t)) = (\bar{x}_{s'}(t) - \bar{\mu}_{s'})^2 - v_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) \qquad (2.2)$$

Given the unbiased estimator, $\hat{B}$, of the bias of $\bar{x}_s(t)$, an unbiased estimator of the true proportion, $\bar{\mu}$, is given by

$$\hat{\bar{\mu}} = \bar{x}_s(t) - \hat{B}$$

$$= \bar{x}_s(t) - (\bar{x}_{s'}(t) - \bar{\mu}_{s'}),$$

with $$V(\hat{\bar{\mu}}) = E_t E_s V_{s'} (\hat{\bar{\mu}}) + E_t V_s E_{s'} (\hat{\bar{\mu}}) + V_t E_s E_{s'} (\hat{\bar{\mu}})$$

$$= E_t E_s V_{s'} (\bar{x}_{s'}(t) - \bar{\mu}_{s'}) + E_t V_s E_{s'} (\bar{\mu}_{s'})$$

since $E_{s'} (\hat{\bar{\mu}}|t,s) = E_{s'} (\bar{\mu}_{s'}|t,s)$

and $V_t E_s E_{s'} (\hat{\bar{\mu}}) = 0$

$$\therefore\ V(\hat{\bar{\mu}}) = E_t E_s V_{s'} (\bar{x}_{s'}(t) - \bar{\mu}_{s'}) + E_t V_{ss'} (\bar{\mu}_{s'}) - E_t E_s V_{s'} (\bar{\mu}_{s'})$$

If $v_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'})$ is a variance estimator such that

$$E\{v_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) \mid t,s\} = V_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'}),$$

if $v_{ss'}(\bar{\mu}_{s'})$ is a variance estimator such that

$$E[v_{ss'}(\bar{\mu}_{s'}) \mid t] = V_{ss'}(\bar{\mu}_{s'})$$

and if $v_{s'}(\bar{\mu}_{s'})$ is a variance estimator such that

$$E[v_{s'}(\bar{\mu}_{s'}) \mid s,t] = V_{s'}(\bar{\mu}_{s'})$$

then an unbiased estimator of $V(\hat{\bar{\mu}})$ is given by

$$\hat{V}(\hat{\bar{\mu}}) = v_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) - v_{s'}(\bar{\mu}_{s'}) + v_{ss'}(\bar{\mu}_{s'})$$

In the case where the original survey is a Census, an unbiased estimator, $\hat{B}$, of the bias of $\bar{X}(t)$ is again given by

$$\hat{B} = (\bar{x}_{s'}(t) - \bar{\mu}_{s'})$$

and an unbiased estimator of the true proportion is given by

$$\hat{\bar{\mu}} = \bar{X}(t) - (\bar{x}_{s'}(t) - \bar{\mu}_{s'})$$

In the expression for the variance of $\hat{\bar{\mu}}$ derived above for the sample survey case, s now becomes the total population. The second and third terms therefore cancel and we get

$$V(\hat{\bar{\mu}}) = E\, V\, (\bar{x}_{s'}(t) - \bar{\mu}_{s'})$$

and therefore an unbiased estimator of $V(\hat{\bar{\mu}})$ is given by

$$V(\hat{\bar{\mu}}) = v_{s'}[\bar{x}_{s'}(t) - \bar{\mu}_{s'}]$$

2.3 Example

The purpose of this section is to describe how the previous estimators have been applied to a small study that was carried out in connection with the 1971 Census, and to present some numerical results.

When the 1971 Census data on type of dwelling were obtained it was suspected that certain categories, namely apartments and duplexes, had been grossly under-reported. As a result, a series of small scale studies were undertaken in an attempt to identify the sources of error.

One of these studies was carried out in the Ottawa region. One of its objectives was to compare the respondents' answers to the type of dwelling question in the 1971 Census with the 'true' type of dwelling as determined by visual observation by an expert. This comparison was carried out on all households in twelve Enumeration Areas (EA's).

Since this study fits the framework developed in the previous section, the sample data were taken to illustrate the use of these estimators in a particular application. It should be noted however, that the figures presented are subject to fairly high sampling variability since they are based on a very small cluster sample. The specific estimators used will now be described.

Suppose that the total population is divided into K enumeration assignments. Let M_k be the size of the kth EA and let $\bar{M}$ be the average size of the EA's. Now suppose a simple random sample of k_o EA's is selected from the total, K, and that within each EA all units are observed. Let this sample be denoted by s' and suppose that true values are determined for each unit in s'.

Define

$x_{ik}(t)$: observed value of unit i in EA k at trial t

μ_{ik}: true value of unit i in EA k

$e_{ik}(t) = x_{ik}(t) - \mu_{ik}$: response deviation of unit i in EA k at trial t

Then,

$$\bar{x}_{s'}(t)^* = \frac{1}{k_o \bar{M}} \sum_{k \in s'} \sum_{i} x_{ik}(t), \quad \bar{\mu}_{s'} = \frac{1}{k_o \bar{M}} \sum_{k \in s'} \sum_{i} \mu_{ik}$$

$$\bar{e}(t) = (\bar{x}_{s'}(t) - \bar{\mu}_{s'})$$

Also,

$$\bar{x}_k(t) = \frac{1}{M_k} \sum_{i=1}^{M_k} x_{ik}(t), \quad \bar{\mu}_k = \frac{1}{M_k} \sum_{i=1}^{M_k} \mu_{ik}$$

An unbiased estimator of the sampling variance $V_{s'}[\bar{x}_{s'}(t) - \bar{\mu}_{s'}]$ for this sample design is given by

$$v_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) = v_{s'}(\bar{e}(t))$$

$$= \frac{K-k_o}{K-k_o} \cdot \frac{1}{k_o-1} \sum_{k=1}^{k_o} \left[\sum_{i=1}^{M_k} \frac{e_{ik}(t)}{\bar{M}} - \bar{e}(t) \right]^2$$

$$= \frac{K-k_o}{K(k_o-1)} \cdot \frac{1}{k_o \bar{M}^2} \left\{ \sum_{k=1}^{k_o} \left(\sum_{i=1}^{M_k} e_{ik}(t) \right)^2 - k_o \bar{M}^2 \bar{e}^2(t) \right\}$$

Substituting this expression into the MSE formula given in equation 2 gives the following unbiased estimator of MSE $[\bar{x}(t)]$.

---

* This estimator was used because:
1) it is unbiased and thus corresponds to $\bar{x}_{s'}(t)$ in the estimation theory
2) the EA's do not vary very much in size

$$\hat{MSE}[\bar{x}(t)] = \frac{1}{K(k_o-1)} \{k_o(K-1)(\bar{x}_{s'}(t)-\bar{\mu}_{s'})^2 - \frac{(K-k_o)}{\bar{M}^2 k_o} \sum_{k=1}^{k_o} M_k^2(\bar{x}_k(t)-\bar{\mu}_k)^2\} \quad (2.3)$$

If, as frequently occurs, $\bar{X}(t)$ represents the Census proportion of the total population in a given category, then both $x_{ik}(t)$ and μ_{ik} are 0-1 variables. The sample for EA k can therefore be split according to the following table of frequencies.

Table 1:

'True' classification μ_{ik}	Census Classification $x_{ik}(t)$ 1	0	TOTAL
1	a_k	b_k	a_k+b_k
0	c_k	d_k	c_k+d_k
TOTAL	a_k+c_k	b_k+d_k	M_k

A term often used to measure errors in Census classification is the net difference rate, which, for EA k, is defined as follows

$$r_k = \frac{c_k-b_k}{M_k}$$

For the total sample we define the net difference rate r as

$$r = \frac{c-b}{k_o\bar{M}} \quad \text{where } c = \sum_{k=1}^{k_o} c_k, \quad b = \sum_{k=1}^{k_o} b_k$$

The quantity r is identical to $\bar{e}(t)$ and thus provides an unbiased estimate of the bias in the Census statistic, $\bar{X}(t)$.

In terms of r and r_k, equation (3) may be expressed as

$$\hat{MSE}\ [x(t)] = \frac{1}{K(k_o-1)}\ \{k_o(K-1)\ r^2 - \frac{(K-k_o)}{\bar{M}^2 k_o}\ \sum_{k=1}^{k_o} M_k^2\ r_k^2\}$$

If we assume that k_o is large and that $k_o << K$ this expression may be further simplified to

$$\hat{MSE}\ [\bar{x}(t)] = r^2 - \frac{1}{(\bar{M}\ k_o)^2}\ \sum_{k=1}^{k_o} M_k^2\ r_k^2$$

Similarly, the sampling variance $v_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'})$ may be expressed in terms of r and r_k and simplified to

$$v_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) = \frac{1}{(\bar{M}\ k_o)^2}\ \sum_{k=1}^{k_o} M_k^2\ r_k^2 - \frac{r^2}{k_o}$$

These two expressions can be easily calculated from the above table of frequencies.

The numerical results are summarized in Table 2. These results confirm the original hypothesis that apartments and duplexes were under-estimated in the 1971 Census.

Table 2: Measures of Reliability of Census Statistics on Type of Dwelling

Type of Dwelling	Census[1] Percentage $\bar{x}(t)\times 100\%$	RMSE $(\bar{x}(t))$	Estimated Bias $(\bar{x}(t))$	Standard Error of Estimated Bias $(\bar{x}(t))$	Estimated True Percentage	Standard Error[2] of Estimated True Percentage
	%	%	%	%	%	%
Single detached.	46.4	1.38	1.69	0.98	44.7	0.98
Double	7.6	2.08	2.20	0.70	5.4	0.70
Duplex	8.1	2.06	-2.36	1.17	10.5	1.17
Single attached and row	8.3	0.44	0.62	0.45	7.7	0.45
Apartment	29.2	2.26	-2.42	0.88	31.6	0.88

[1]Total population includes 0.4% mobile dwellings.

[2]Equivalent to the standard error of the bias.

2. APPROACH II

Under this approach we use as a measure of error the deviation of the specific Census figure from the true parameter being estimated. This differs from the usual response error model in as much as no probability model is assumed on observations being made at Census.

Let $\bar{\mu}$ be the population mean being estimated and let $\bar{x}$ be the corresponding Census figure. Then the net error involved in using $\bar{x}$ as an estimate of $\bar{\mu}$ is given by

$$E = \bar{x} - \bar{\mu}$$

Assume that for a sample s' of the population, the true values can be determined. Let $\bar{x}_{s'}$ and $\bar{\mu}_{s'}$ be unbiased estimates of $\bar{x}$ and $\bar{\mu}$ obtained from the sample respectively. Then an unbiased estimator of E is given by

$$\hat{E} = \bar{x}_{s'} - \bar{\mu}_{s'}$$

The above estimate can be used to produce an alternate estimator of $\bar{\mu}$ by correcting $\bar{x}$ as follows:

$$\hat{\bar{\mu}} = \bar{x} - (\bar{x}_{s'} - \bar{\mu}_{s'})$$

for which the sampling variance is given by

$$V(\hat{\bar{\mu}}) = V(\bar{x}_{s'} - \bar{\mu}_{s'})$$

It is easy to show that $\hat{\bar{\mu}}$ will be more reliable than $\bar{\mu}_s$ when $\bar{x}_{s'}$ and $\bar{\mu}_{s'}$ are highly correlated, which is usually expected.

The above can also be extended to sample surveys when true values are known for a sub-sample of the original sample.

4. CONCLUSION

This paper has presented two approaches for measuring the reliability of Census and sample survey data when true values can be determined for a sample or sub-sample of the population. For each approach, a method of correcting the original estimate was also presented.

Although this theory was developed mainly for applications to response error problems in Census, it is also applicable to other types of situations, e.g. coverage errors, coding errors etc. The particular approach would therefore in general depend on the type of error being investigated.

RESUME

Madow (1968) a proposé un schéma d'échantillonnage à deux degrés suivant lequel le biais de réponse peut être éliminé des enquêtes par sondage en obtenant des valeurs "réelles" pour un sous-échantillon de l'échantillon original. Comme c'est souvent le cas aux recensements ou aux enquêtes en cours, les données des sous-échantillons ne servent pas à corriger les estimations de l'enquête principale, mais à évaluer leur fiabilité. Ce document vise d'abord à présenter des méthodes permettant d'obtenir des estimations de fiabilité lorsque les valeurs "réelles" peuvent être établies pour un sous-échantillon d'unités.

REFERENCES

[1] Madow (1965), "On Some Aspects of Response Error Measurement", Proceedings of the ASA, Social Statistics Section, pages 182-192.