ESTIMATION OF INTERVIEWER VARIANCE FOR CATEGORICAL VARIABLES

S. Lynne Stokes, University of Texas at Austin
Mary H. Mulry-Liggan, Bureau of the Census

1. Introduction

Errors introduced in the measuring, editing, or coding of responses in a sample survey affect the behavior of the estimators obtained from the sample and sometimes affects our ability to measure that behavior. Models designed to measure the impact of these errors indicate that the non-sampling errors may contribute substantially to the bias and/or variance of the estimators obtained from the sample. Furthermore, when these errors are positively correlated within the sample, as they might be when a single operator, such as an interviewer or coder, handles a number of cases, the usual estimators of the standard errors of means and totals are likely to be biased downward. This bias is called the correlated component of response variance. If good estimates of the correlated component can be made, the estimates of the standard errors can be improved and problem items can be identified.

Most methods for estimating the correlated component require interpenetration of operators, a technique introduced by Mahalanobis [9]. In its most basic form, interpenetration requires the random sample of size n from a population of size N to be randomly divided into k subsamples of size m = n/k, and each subsample to be assigned to a single operator. Then the typical model describing y_{ijt}, the recorded value in the t^{th} survey replication for unit j, which is in operator i's assignment, is

$$y_{ijt} = \mu_j + e_{ijt}, \tag{1.1}$$

where $\mu_j = E(y_{ijt} | j)$ and e_{ijt} is the error in that recorded value. Then for $\bar{y} = \Sigma\Sigma y_{ijt}/km$, we have

$$V(\bar{y}) = \frac{1}{km} V(\mu_j + e_{ijt}) + \frac{1}{k}\frac{m-1}{m} Cov(e_{ijt}, e_{ij't}) \tag{1.2}$$

if $Cov(e_{ijt}, e_{i'j't}) = Cov(\mu_j, e_{ijt}) = 0$.

We will refer to the operator introducing the correlated error as an interviewer, since that is a common source for such errors.

We examine a model proposed by Kish [7] for the correlated errors in continuous data. He decomposed the error term in (1.1) as $e_{ijt} = b_i + e'_{ijt}$, where b_i can be thought of as a random variable associated with the i^{th} interviewer and represents the average bias that she introduces into a measurement. Then e'_{ijt} represents the composite of all other uncorrelated non-sampling errors (i.e., only one source of correlated error is assumed). Then (1.1) can be rewritten as

$$y_{ijt} = \mu + b_i + e_{ijt}, \tag{1.3}$$

where $\mu = E_j \mu_j$ and $\epsilon_{ijt} = (\mu_j - \mu) + e'_{ijt}$ contains both sampling and uncorrelated non-sampling errors. Then (1.2) can be written as

$$V(y) = \frac{1}{km} V(y_{ijt})[1 + (m-1)\rho_b]$$

where $\rho_b = V(b_i)/V(y_{ijt}) = \sigma_b^2/(\sigma_b^2 + \sigma_e^2)$, if we assume that $Cov(b_i, b_j) = Cov(b_i, \mu_{ijt}) = 0$. Then the correlated component $CC = \sigma_b^2$.

Biemer and Bailar [1] model the mechanism causing the correlated errors differently when the characteristic being observed is membership in a category. For each category, an interviewer can make two types of errors: ϕ_i is the probability that interviewer i records a unit reporting that it belongs to the category as not belonging to it and θ_i is the opposite kind of error. Then (ϕ_i, θ_i) is considered to be a random vector associated with the i^{th} interviewer. We can avoid consideration of individual characteristics of ϕ_i and θ_i by defining a new random variable $p_i = E(y_{ijt} | i)$, the probability that interviewer i records a randomly chosen unit from a random replicate as belonging to the category. Then the categorical data equivalent of (1.3) is

$$y_{ijt} = Ep_i + (p_i - Ep_i) + \epsilon_{ijt}, \tag{1.4}$$

where $\epsilon_{ijt} = (\pi_j - Ep_i) + e'_{ijt}$, $\pi_j = E(y_{ijt} | j)$.

Then the usual estimator for proportion, $\hat{P}$, has variance

$$V(\hat{P}) = \frac{1}{km} V(y_{ijt})[1 + (m-1)\rho_p] \tag{1.5}$$

where $\rho_p = V(p_i)/V(y_{ijt}) = \sigma_p^2/\mu_p(1-\mu_p)$, with $\mu_p = Ep_i$. Then the correlated component $CC = \sigma_p^2$.

2. The Estimator

Let us denote the m units in interviewer i's assignment by y_{ij}, j = 1, ..., m. We suppress the subscript t since in this paper only one survey replicate is assumed. (This simply means that we are unable to distinguish response variance from non-sampling variance.) Then the usual ANOVA estimator for the variance component of random model,

$$CC = \frac{S_b^2 - S_w^2}{m} \tag{2.1}$$

where $S_b^2 = \frac{m}{k-1} \sum_i (\bar{y}_{i\cdot} - \bar{\bar{y}})^2$

and $S_w^2 = \frac{1}{k(m-1)} \Sigma\Sigma_{ij} (y_{ij} - \bar{y}_{i\cdot})^2$

is an unbiased estimator of CC for both models under the assumptions stated in the previous section.

The precision of $\hat{CC}$ is needed for planning of interpenetration studies, but it has not been investigated thoroughly by analytical methods. The variance of $\hat{CC}$ when both ϵ_{ij} and b_i of (1.3) are normally distributed is given in Searle [10] for a simple random sample. MacLeod and Krotki [8] reported the variance of Fellegi's [6] more complex estimator of the correlated component by empirical variance estimation techniques. Biemer and Stokes [2] provided analytical expressions for the variance of Fellegi's estimator under the restrictive assumption of normality of both the error and interviewer bias terms. However, these assumptions are inappropriate for the categorical error model given by (1.4). In this paper, we find out how much can be learned about the variance of $\hat{CC}$ for the discrete model while making as few assumptions as possible about the distribution of p_i.

3. Variance of the Estimator

The variance of $\hat{CC}$ as given by (2.1) under model (1.3), adding the assumption of normality of the ϵ_{ij}'s is

$$V(\hat{C}\hat{C}) = \frac{1}{k}\left[\mu_{b4} - \frac{k-3}{k-1}\sigma_b^4\right] + \frac{4}{(k-1)m}\sigma_\varepsilon^2\sigma_b^2 + \frac{2}{m^2(k-1)}\sigma_\varepsilon^4 + 0\left(\frac{1}{m^3}\right) \quad (3.1)$$

where $\mu_{b4} = Eb_i^4$. The first term of this expression would be the variance of an estimate of σ_b^2 if the b_i values were known. Because they must be estimated, however, $V(\hat{C}\hat{C})$ is increased by the 2nd and 3rd terms. If we add the assumption of normality of the b_i's, we obtain from (3.1) the well-known expression

$$V(\hat{C}\hat{C}) = \frac{2}{k-1}\left[\sigma_b^4 + 2\sigma_b^2\frac{\sigma_\varepsilon^2}{m} + \frac{\sigma_\varepsilon^4}{m^2} + 0\left(\frac{1}{m^3}\right)\right]. \quad (3.2)$$

For the categorical model given by (1.4), one can show, after much tedious algebra, that

$$V(\hat{C}\hat{C}) = \frac{1}{k}\left[\mu'_{p4} - \frac{k-3}{k-1}\sigma_p^4\right] + \frac{4}{km}\left[-\mu_{p4} + \mu_{p3}(1+2\mu_p) - \mu_{p2}\mu_p(2+\mu_p) + \mu_p^3 + \frac{1}{(k-1)}\cdot\sigma_p^2(\mu_p - \mu_{p2})\right] + \frac{2}{km^2}\left[3\mu_{p4} - \mu_{p3}(5+2\mu_p) + \mu_{p2}(3\mu_p+2) - \mu_p^2 + \frac{1}{k-1}(\mu_{p2}+\mu_p)^2\right] + 0\left(\frac{1}{m^3}\right) \quad (3.3)$$

where $\mu_{pr} = Ep_i^r$ and $\mu'_{p4} = E(p_i - \mu_p)^4$.

When planning an interpenetration experiment, one goal might be to determine the sampling plan required to achieve a specified coefficient of variation. If the assumption of normality of the b_i's were acceptable, then we would find from (3.2) that

$$(CV)^2 = \frac{2}{k-1}\left[1 + \frac{2}{m}\left(\frac{1-\rho_b}{\rho_b}\right) + \frac{1}{m^2}\left(\frac{1-\rho_b}{\rho_b}\right)^2 + 0\left(\frac{1}{m^3}\right)\right] \quad (3.4)$$

Kish [7] gives ranges for the size of ρ for different types of questions. They range from 0 for factual questions to about .10 for subjective or difficult ones. For a specifically assumed ρ, required values of m and k can be determined from (3.4) to achieve a desired $(CV)^2$. The appearance of μ_{b4} in (3.1) prevents a similar calculation from being possible if a distribution for b_i is not assumed.

The picture is even worse for the categorical case. Notice that all terms of (3.3) (rather than just the first as in (3.1)) involve moments of p_i higher than the 2nd. Without knowing something about the magnitude of these moments relative to σ_p^2, we can't determine the sampling plan required for meeting our CV requirements. In this paper, our aim is to see how much we do know about the 3rd and 4th moments of p_i and thus about $V(\hat{C}\hat{C})$ for the categorical model, while making as few assumptions about the distribution of p_i as possible. We use this information to find bounds on sample sizes needed to achieve specified CV levels.

4. Variance Bounds in the Categorical Model

Our goal is to obtain bounds for the terms of (3.3) for specified survey design parameters k and m. If we can obtain bounds that are sufficiently narrow, they might be used for planning an interpenetration study to achieve a desired CV, for example, as was suggested in the previous section. In addition, we might be able to tell, with the help of such bounds, how much the erroneous use of the continuous model formulae (such as (3.2) and (3.4)) for planning experiments about categorical variables can mislead us.

Unfortunately, unless we include some restrictions on the distribution of p_i, $V(\hat{C}\hat{C})$ is so variable that no useful general conclusions of the type we wish to draw can be made. So we arbitrarily include the assumption that the 3rd central moment of p_i is 0. This assumption is weaker, of course, than that of normality of the b_i's in the continuous model.

The bounds for $V(\hat{C}\hat{C})$ in the categorical model are found by using a corollary (DeVylder [5], Brockett [3]) of the Markov-Krein theorem, which provides the best upper and lower bounds on the expected value of certain functions of a bounded random variable whose first three moments are known. The corollary, tailored to our application, is as follows:

Let X be a random variable having range [0,1]

with $EX = \mu$, $V(X) = \sigma^2$, and $\mu_3 = E(X-\mu)^3$

known. Then for any h for which $h^{(4)}(x) \geq 0$,

$$h(c_1)\pi_1 + h(c_2)(1-\pi_1) \leq Eh(X) \leq h(0)\eta_1 + h(d)\eta_2 + h(1)(1-\eta_1-\eta_2) \quad (4.1)$$

where

$$d = \frac{\mu_3 - (1-2\mu)\sigma^2}{\sigma^2 - \mu(1-\mu)} + \mu,$$

$$\eta_1 = \frac{\sigma^2 + (d-\mu)(1-\mu)}{d} \qquad \eta_2 = \frac{\mu(1-\mu)-\sigma^2}{(1-d)d}$$

$$c_1 = \frac{\mu_3 - \mu_3^2 + 4\sigma^6}{2\sigma^2} + \mu \qquad c_2 = \frac{\mu_3 + \mu_3^2 + 4\sigma^6}{2\sigma^2} + \mu$$

$$\pi = 1/2 + \frac{\mu_3}{\mu_3^2 + 4\sigma^6}$$

Since the terms of $V(\hat{C}\hat{C})$ for the categorical model can be expressed as $Eh(X)$ for an appropriate h (or perhaps as $Eh_1(X) - Eh_2(X)$ if the condition concerning the non-negative derivative is not met), we may use (4.1) to find upper and lower bounds for $V(\hat{C}\hat{C})$ for a characteristic with a specified μ_p and ρ_p.

Kish's ranges can help us choose a suitable value of ρ_p for a certain type of question and μ_p is the expected proportion recorded in the category. Then we see from (1.5) that $\sigma_p^2 = \mu_p(1-\mu_p)\rho_p$. Table 1 illustrates the results of this procedure by displaying bounds for $V(\hat{C}\hat{C})$ where $\rho_p = .1$, $k = 2$, and 2 values of interviewer workloads, m, correspond to a telephone survey (m=25) or a census (m=500).

There is an analogous corollary (DeVylder [5], Brockett [3]) of the Markov-Krein theorem that provides the best bounds for E(h(X)) when only the first two moments of a bounded random variable are known. Unfortunately, as mentioned earlier, the bounds achieved without the assumption on the third central moment are too wide to be of much value for our purposes.

An important point to emphasize is that these corollaries of the Markov-Krein theorem yield "tight" bounds in the sense that they can not be any shorter. The corollaries actually produce distributions which satisfy the assumptions and achieve the bounds.

5. Application to Sampling Designs

With the bounds obtained for $V(\hat{CC})$, we can address obtaining the optimal design for an experiment to estimate CC. Two criteria for determining the number of interviewers to interpenetrate are considered: cost and coefficient of variation.

The cost of conducting a personal interview survey with k interviewer assignments interpenetrated is modeled by $C = C_0 + C_1 \text{ km}\sqrt{k}$ where C_0 is overhead cost of the survey and C_1 is the cost of each interview when the interviewer assignments are not interpenetrated. The model is based on the assumption that the increased cost arises from the increase in travel expenses (Cochran [4]). The assumption that the cost increases by a factor $\sqrt{k}$ is based on the fact that the average distance between randomly distributed points in a plane is increased by $\sqrt{k}$ when the density of those points is decreased by a factor of k. Numerical studies indicated that for fixed cost and total sample size n = mk, $V(\hat{CC})$ is a decreasing function of k for both the continuous and categorical models. Therefore, k should be chosen as large as resources will allow in each case. So using the continuous model to make such a decision about categorical variables (as is often done) leads to the right allocation of resources.

A model for the cost of interpentrating k interviewer assignments in a telephone survey has not been developed.

The coefficient of variation appears to be a more influential criterion in the determination of the interpenetration requirements than the cost. We determined numerically the required number of interviewers, each having assignment size m, to achieve a specified coefficient of variation. This approach was taken since the survey designer knows the number of interviews a full-time interviewer can be expected to complete during the survey period. A previously mentioned, m=500 is appropriate for a census interviewer, while m=25 is more reasonable for a two-week telephone survey.

Table 2 shows the results that were obtained for the two models. The range of k, number of interviewers required to achieve a coefficient of variation (CV) of 0.5 when $\rho_p = .1$ and m is fixed is given for several values of ρ_p for the categorical model. The last row of the table shows the number of interviewers required to achieve CV = .5 when $\mu_b = .1$ for the continuous model where the b_i's are normally distributed. (This can easily be shown from from (3.4)). The intervals yielded by (4.1) are quite short when μ_p is near 0. As μ_p nears .5, there is a wider range of distributions which can satisfy the moment conditions and thus (4.1) cannot pinpoint the required k as well. Note that when μ_p is close to 0, using (3.2) as a proxy for the correct variance formula for choosing the interpenetration sample design is overly pessimistic; i.e., even in the worst possible case, a better CV than expected would be achieved if the normal theory required k's were used. There are distributions of p_i for larger μ_p, however, for which choosing the k suggested by the continuous model would yield unacceptably low precision.

A surprising observation from the table is that the values for m=25 and m=500 are so similar. Shouldn't we expect a much smaller k when each interviewer's workload is 20 times as large? Certainly the required k to achieve a fixed CV will always be smaller when m=500 than when m=25, but it may not be by much, if the largest component of $V(\hat{CC})$ is the first term of (3.3) (or of (3.1) or (3.2), since the same holds for the continuous model), which happens when ρ is large. The first term of (3.3) doesn't involve m, and that is where most of the uncertainty from not being able to pinpoint the distribution of p_i does its damage. When μ_p is closer to 0, the ranges for k yielded by (4.1) are wider, and the intervals for m=25 and m=500 are more widely separated.

In general, the bounds widen rapidly as the CV decreases. Both the upper and lower bounds increase with increasing CV, as would be expected.

It might seem natural to compare the classical continuous model with the categorical one in which p_i behaves like a normal random variable in the sense that its first 4 moments match that of a normal. (It obviously cannot be exactly normal since it is bounded on (0,1).) For a large portion of the square making up the range of interest $(0 \le \mu_p \le 1,\ 0 \le \rho_p \le .1)$ there doesn't exist a bounded random variable having its first 4 moments match those of a normal. So the natural comparison isn't possible. For those that were possible, we found still a large discrepancy, in some cases, between the required k's for the two models.

For both the categorical and continuous model, for any specified ρ, increasing m beyond a certain point is not helpful in improving the precision. This can be done only by increasing the number of interviewers. The increased number of interviewers, it turns out, improve the precision whether they are interpenetrated all together or in smaller groups. This information suggests, for example, that when several surveys are being run out of the same telephone facility, it would be helpful for estimating CC to have each interviewer working on more than one survey, so that a larger number of interviewers can complete interviews for each.

TABLE 1

Bounds of V(CC) under the categorical model with $\rho_p = .1$ and k = 2

p	m=25	m=500
	$\times 10^{-3}$	$\times 10^{-4}$
.10	(.23, .28)	(.87, 1.3)
.30	(.90, 3.0)	(4.7, 23.)
.50	(1.2, 4.5)	(6.5, 35.)

TABLE 2

Ranges of Required k to Achieve CV=.5 when $\rho_p = .10$

p	m=500	m=25
.10	(3.5, 6.2)	(10.0, 13.3)
.20	(3.5, 26.4)	(9.0, 29.7)
.30	(3.5, 32.9)	(8.7, 34.9)
.40	(3.5, 35.5)	(8.6, 37.0)
.50	(3.5, 36.2)	(8.5, 37.6)
normal theory	9.3	15.8

REFERENCES

1 Biemer, P.P., and Bailar, Barbara A. (1981). Some Methods for Evaluating Nonsampling Error in Household Censuses and Surveys, presented at the Conference of European Statisticians in Geneva, Switzerland in June 1981.

2 Biemer, P.P., and Stokes, S.L. (1984). Optimal design of interviewer variance experiments in complex surveys, to appear in JASA.

3 Brockett, Patrick L. and Cox, Samuel H. (1984). Insurance calculations using incomplete information. Working Paper 83/84-2-22, Department of Finance, The University of Texas at Austin.

4 Cochran, William G. (1977). Sampling Techniques, 3rd Edition. John Wiley and Sons, New York.
5 DeVylder, F. (1983). Maximization, under equality constraints, of a functional of a probability distribution. Insurance: Mathematics and Economics, Vol. 2, 1-16.
6 Fellegi, I.P. (1974). An improved method of estimating the correlated response variance, JASA, 69, 469-501.
7 Kish, Leslie (1962). Studies of interviewer variance for attitudinal Variables. JASA, 57, 92-115.
8 MacLeod, A. and Krotki, K.P. (1979). An empirical investigation of an improved method of measuring correlated response variance. Survey Methodology, 5, No. 2, 59-78.
9 Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. JASA, 109, 325-370.
10 Searle, S.R. (1971). Linear Models, John Wiley and Sons, New York.