Variance Components and Design Comparisons for an Area Sample Survey in Cameroon

C. H. Proctor
Department of Statistics, North Carolina State University
Raleigh, NC 27695-8203

In this paper we will discuss uncertainties in estimates produced by a survey of root crop production and consumption in Cameroon. Our main purpose is to improve the design and conduct of future such surveys. This survey has been completed and the standard errors of its estimates are set. For most of its information it is the only source, and users are well advised to make the most of it. It stands as a valuable and unique accomplishment.

There are two major sources of uncertainty in surveys. One is sampling and the other is measurement. Sampling uncertainty in the present survey arises from having gone to only 64 of the 43,376 area segments. By having gone to all 43,376, and thereby having visited every household in the area, one could have brought sampling uncertainty to zero. Alternatively, if every sampling unit had been just like every other one then a sample of size one would have had zero uncertainty. That is, sampling uncertainty depends both on variability among the units as well as on sample size.

Measurement uncertainty is a bit more complicated matter. For most items of information there is a physical reality that the questionnaire wording and enumerators' questions are aimed at revealing. For example, the amount of taro produced at a given household during the year is a physically well-determined weight, but the amount written on the form depends on the respondent's experiences with the production (present or not at harvests, for example), memory factors, willingness to be truthful, and on the enumerator's explanation of the quantity required (amount harvested by "just these people" or by "this larger group"). For this survey in Cameroon the problems of language differences between enumerator and respondent and of keeping constantly in mind the exact geographic limits of the selected segment could easily have led to measurement uncertainty.

Measurement uncertainty can be studied by comparing duplicate determinations. That is, one enumerator with a given questionnaire visits an area segment and then next week another enumerator with another form visits the same segment again. If these visits are continued over several weeks and done at a number of area segments one can compare the recorded amounts and begin to learn what form of questioning and which enumerators seem to yield stable responses. Obviously, such a study is too costly in most cases. However, just from our survey results we can gain some idea of the variability introduced by the enumerators, but we cannot know how much was caused by the questionnaire, since only the one form was used.

Amounts of variability are more precisely expressed by variance components. The survey data will be used to furnish estimates of variability from segment to segment and from enumerator to enumerator. Once we have these estimates, we can begin to see how changes in the survey design would

affect standard errors of the survey estimates. Variance components are oftentimes estimated by doing analyses of variance and we will now describe the analyses of variance that we used.

The area segments assigned to any one survey team will be called a team-sector. Some resemblance between team-sectors and provinces was built into the design but the resemblance is not perfect. Within each team-sector there are a number of (paper) zones, defined in the sampling plan, each containing four sampled area segments. The zones are akin to geographic slices of a team-sector. The four sampled area segments of a zone belong to each of the four interpenetrating subsamples. Each team had three enumerators. Finally, within each area segment there were more or less 10 households.

For the analyses of variance there are thus five factors: Team-sector (T), Zone (Z), Area segment (S), Enumerator (E) and Household (H). In conventional ANOVA notation the following seven sources of variation are specified for the analyses as:

Source	Notation	DF
1. Team-sectors	T	4
2. Zones in Team-sector	Z(T)	24
3. Segments in Zone	S(Z*T)	38
4. Enumerators in Team-sector	E(T)	12
5. Enumerators by Zones in Team-sector	E*Z(T)	43
6. Enumerators by Segments in Zone	E*S(Z*T)	72
7. Households in Enumerator by Segment	Error	441

The structure of the sources seems reasonably clear but the degrees of freedom (DF as discovered by PROC GLM) quantities show there is unbalance in the design. A survey is not an experiment. For example, enumerators who were working in Douala City ended by doing some interviewing in the rural area, and thus five team-sectors were defined rather than four.

The first two sources, Teams and Zones, will be taken as fixed. In a geographic sense (as Sector) they are, but Teams in a measurement sense are but a sample of those which we could have formed. However, we cannot separate the possible geographic (fixed) differences among sectors from the possible (random) biases induced by the teams. The third source contains sampling variability and so does the seventh. The fourth, fifth and sixth sources are measurement variabilities.

The survey design is, of course, unbalanced as the degrees of freedom indicate. It is possible to construct a hypothetical balanced design that has degrees of freedom close to those recorded. In this design we put 5 Team-sectors, 5.8 Zones per Team-sector, 2.31 segments per zone, 3.4 enumerators per team and 2.95 households per enumerator assignment in a segment. Using these (noninteger) numbers of levels one can compute coefficients of variance components in expected mean squares. We then ran PROC GLM to get the mean squares and used a program called LSVC (Least squares for variance components) to compute estimates of the variance components (see Proctor, 1985). The results appear in Table 1.

Also shown in Table 1 are

estimates based on a likelihood maximization calculation. This was done from PROC MIXMOD, a SAS procedure (see Giesbrecht, 1983). The ML estimators take into account the unbalance in the design and thus could be somewhat better if the distributions are normal. An introductory discussion of variance components is found in Chapter 10 of the Snedecor and Cochran (1967) textbook.

A small scale simulation using this survey's configuration of levels and unbalance was done to compare three estimators of variance components, ML and MML of PROC MIXMOD and LSVC, the ANOVA-estimate. We found that the ML estimates tended to go negative quite often even when the true component was over 10% of error variance. On the other hand the ML estimates were more sensitive to the relative sizes of the components than were the LSVC or the MML estimates. As a compromise estimator we suggest setting the zero values equal to 10% of error variance and then averaging ML with LSVC. This opportunistic or "seat-of-the-pants" estimator was in fact not computed and variances were projected separately as will be seen.

For the amounts consumed, the enumerator sources are large (20% or 30%), while the segment variabilities are small (0% to 10%). For the amounts produced, both sources are considerable, segment variability at 15% or 25% and enumerator variability at 20% or 10%. For the demographic indices, segment to segment variability is low (except for adult schooling level) and so is enumerator variability.

Since we did not repeat any interviews at any household we have no way of telling exactly how much of household variance is measurement error and how much is actual (sampling) variation among households. The percentages of enumerator variability in Table 1 can be converted to rough reliabilities by dividing by 100 and subtracting from 1. That is, reliabilities of the consumption amounts appear to be about .70 or .75. One would expect that these same reliabilities would apply to household variance.

Table 1. Variance Component Estimates for Square Root Transformed Amounts Consumed of Four Foods, for Four Demographic Variables and for Two Crop Amounts Produced in Logarithms

Source	Estimation Method:	Cocoyam LSVC	Cocoyam ML	Yam LSVC	Yam ML	Cassava LSVC	Cassava ML	Sweet Potato LSVC	Sweet Potato ML
3. Segments (Team-Zone)		.44	.01	.59	.06	.30	.03	.25	0
4. Enumerators (Team)		.26	.18	.21	.16	.47	.38	.08	.06
5. Enumerators by Zone (Team)		0	0	.41	0	.25	0	.35	0
6. Enumerator by Segment (Team*Zone)		.81	.70	.15	.51	.48	.59	.43	.57
7. Households (E*S*T*Z)		3.01	3.09	3.78	3.79	2.41	2.44	2.20	2.23
Total Variance		4.52	3.98	5.14	4.52	3.91	3.44	3.31	2.86
Percent Segment Variability		10	0.3	11	1	8	1.2	8	0
Percent Enumerator Variability		24	22.1	15	15	31	28.2	26	22
Mean		3.04		3.56		3.57		2.28	

Table 1. (continued)

Source	No. of Male Adults		No. of Female Adults		No. of Babies		Average Adult School		Log Amount Produced: Cocoyam		Log Amount Produced: Cassava	
	LSVC	ML	LSVC	ML	LSVC	ML	LSVC	ML	LSVC	ML	LSVC	ML
3.	.053	0	.084	0	.021	0	.046	.021	4.10	1.86	2.43	1.26
4.	.032	.014	.072	.036	.021	.006	.002	.002	2.73	1.99	.38	.30
5.	0	0	0	0	.001	0	0	0	.69	0	.60	.29
6.	.085	.078	.299	.236	0	0	.001	0	1.67	1.90	0	.09
7.	.869	.860	1.895	1.874	1.540	1.493	.206	.205	12.31	12.50	4.84	4.83
Total	1.039	.952	2.350	2.146	1.583	1.499	.255	.288	21.50	18.25	8.25	6.77
% Seg.	5	0	4	0	1	0.0	18	10	19	10	29	19
% En.	11	10	16	13	1	0.4	0	1	24	21	12	10
Means	1.31		1.95		1.15		.54		6.24		3.12	

(A "0" signifies the estimate became negative.)

Now let's consider standard error calculations for the design we used, as well as for some variations in the design. For the initial design there were four teams and each was assigned to a sector. We can suppose there were four zones in each sector. There were three enumerators in each team. The fact that supervisors also did a few interviews can be ignored. The variance of a sample mean can be expressed in some generality as:

$$V(\bar{y}) = \sigma_3^2/n_3+\sigma_4^2/n_4+\sigma_5^2/n_5+\sigma_6^2/n_6$$

$$+ r_{xx}\sigma_7^2/n_7+(1-r_{xx})\sigma_7^2/n_7 \ .$$

The σ_i^2 quantities are the variance components and estimates of them come from Table 1. For our initial design $n_3 = 64$, the number of segments, n_4 is the number of enumerators, n_5 the number of enumerator by zone combinations, n_6 the number of enumerator by segment combinations and n_7 is the number of households. The quantity r_{xx} is reliability. For our initial design $n_4 = 12$, $n_5 = 48$, $n_6 = 192$ and $n_7 = 640$, with $r_{xx} = .7$.

For the variance components as estimated for cocoyam consumed we find:

$$V(\bar{y}) = .44/64 + .26/12 + .81/192 + 3.01/640$$

$$= .03746, \text{ with standard error, SE} = .1936 \ .$$

The mean is 3.04 and thus the sampling coefficient of variation is 6% which agrees with earlier findings based on variability among the replicated subsamples. Suppose

Table 2. Variances Projected for Survey Design Variations

Item of Information	Type of Estimation	Original Design (4 teams, 64 SU's)	Eight Teams (64 SU's)	Two Teams (64 SU's)	Two Weeks Training (48 SU's)
Cocoyam Consumed (Square Root)	LSVC	.0375	.0266	.0591	.0367
	ML	.0236	.0161	.0386	.0217
Yam Consumed (Square Root)	LSVC	.0419	.0332	.0594	.0420
	ML	.0228	.0162	.0362	.0218
Cassava Consumed (Square Root)	LSVC	.0553	.0357	.0945	.0505
	ML	.0390	.0232	.0707	.0346
Sweet Potato Consumed (Square Root)	LSVC	.0235	.0202	.0302	.0230
	ML	.0114	.0090	.0165	.0108
No. of Male Adults	LSVC	.0053	.0040	.0080	.0053
	ML	.0029	.0023	.0041	.0029
No. of Female Adults	LSVC	.0118	.0088	.0178	.0116
	ML	.0072	.0057	.0102	.0070
No. of Babies	LSVC	.0045	.0036	.0063	.0048
	ML	.0028	.0026	.0033	.0031
Ave. Adult School	LSVC	.0012	.0011	.0014	.0015
	ML	.0008	.0007	.0010	.0010
Log Cocoyam Produced	LSVC	.3339	.2201	.5614	.3208
	ML	.2243	.1414	.3902	.2109
Log Cassava Produced	LSVC	.0897	.0739	.1214	.0970
	ML	.0587	.0462	.0837	.0618

we had recruited twice as many enumerators (24 instead of 12 in the rural area) and formed 8 teams instead of 4. The eight teams could have been assigned to 8 sectors with two zones in each sector. With the same sample size the $V(\bar{y})$ becomes .02663 with a sampling CV of 5%. Of course, the additional expense would have been considerable although the survey time would have been reduced. In a similar vein, if we had been forced to operate with only two teams the variance would have risen to $V(\bar{y})$ = .05913 with a sampling CV of 8%.

Now we arrive at a bit more complicated and even more hypothetical suggested design. Suppose we had spent 2 weeks in training and then gone to only 3 of the four subsamples. The survey expenses and timing would be roughly the same as the actual design. The additional week of training could have been expected to have improved reliability to some extent. Let's suppose reliabilities could have been improved by 10% to 20%. For example, a reliability of .7 could have been raised to one of .8. We believe this suggestion is conserva tive but it is a supposition. We will represent this by a reduction in the measurement variance component of 15%. If σ^2_{7I} = 3.01 is household variance for the actual survey with .7 reliability then σ^2_{7II} = (.7/.8) 3.01 = 2.63 will be household variance with reliability of .8. The component σ^2_4 = .26 becomes .22 and σ^2_6 = .81 goes to .69.

$$V(\bar{y}) = .44/48 + .22/12 + .69/192 + 2.63/480$$

$$= .03666 \text{ with SE} = .19 \text{ and sample CV} = 6\%.$$

Thus the same uncertainty in the estimate of cocoyam consumed would Thus the same uncertainty in the estimate of cocoyam consumed would have been attained by this modification in design. The results for the other items of information are given in Table 2.

When all items of information in Table 2 are considered we see that the extra week of training plus the reduction in sample size would have been detrimental only for Number of Babies, Average Adult Schooling and Cassava Produced. What we cannot demonstrate by data from just one survey is the amount of overall bias. There is every reason to expect that with the longer training period this survey bias would have been reduced. It should have been possible to arrive at standard suggestions for probes and follow-up questions that would have aided the respondent in better understanding the questionnaire items.

REFERENCES

Giesbrecht, F. (1983). "An efficient procedure for computing minque of variance components and generalized least squares estimates of fixed effects," *Communication in Statistics: Theory and Methods* 12, 2169-2177.

Proctor, C. H. (1985). "Simple inferences for infinite and finite population variance components and for fixed effects variation components," Institute of Statistics Mimeo Series No. 1661, North Carolina State University, Raleigh, NC 27695-8203.

Snedecor, G. W. and Cochran, W. G. (1967). *Statistical Methods.* Sixth Edition, The Iowa State University Press, Ames, Iowa.