Statistics Division
Working Paper Series

NO: ESS / ESSG/010e

Effect of sampling design on reliability of estimates:

The variance of distributions derived from household survey data

April 2007

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS

منظمة الأغذية والزراعة للأمم المتحدة
联合国粮食及农业组织
Food and Agriculture Organization of the United Nations
Organisation des Nations Unies pour l'alimentation et l'agriculture
Organización de las Naciones Unidas para la Agricultura y la Alimentación

The effect of sampling design on the reliability of estimates of the variance of distributions derived from household survey data: a simulation study

Giuseppe Arbia
"G. d'Annunzio" University
Pescara, Italy

On a regular basis, FAO produces estimates of the prevalence of undernourishment and related measures that require estimates of the frequency distribution of household per capita food consumption, expressed in terms of dietary energy (kilocalories). National sample surveys investigating household income and consumption are the usual source of data for these activities. In using the distribution data from such surveys, it is important to ensure that not only the estimate of the mean but also the variance and the shape of the distribution are sufficiently accurate. Inaccurate estimation of these parameters may lead to serious biases in the evaluation of the prevalence of undernourishment. Most of the sample surveys of household income/expenditure/consumption are generally based on a two-stage sample design with stratification of the primary sample units (usually geographical units) and secondary units represented by households. However, apart from the classical results on the estimation of the mean, in the literature there is no clear evidence that the estimation of the other characteristics (such as variance, skew- ness and kurtosis) is unbiased with respect to this kind of design. This paper presents a simulation study that attempts to address this problem in formal terms.

In the simulation, we start with real data taken from the Mauritius 1996-97 Household Budget Survey. We first simulate a virtual population from a log-normal distribution with expected values, variance and sampling frame derived from the real Mauritius survey data. This will allow much more flexibility and generality of results than simply using the real data. However, the results are connected with a real population and are not just artificially laboratory-generated. In a second phase, the virtual population thus generated is sampled by using various sampling strategies, and the results are compared in terms of the distributional properties of the estimates.

The simulation results suggest that the bias of the estimates of the variance is generally very high in all experimental situations considered. In fact, the relative bias ratio always exceeds the threshold value suggested by Cochrane (1963) as a rule of thumb. The bias is emphasized by the presence of intradis-trict correlation (see Table 1). Conversely, it is not significantly affected by different design effects (DEFF) values (see Table 2). Furthermore, the estimates based on two-stage stratified samples are also seriously affected by a low level of efficiency with respect to estimates based on random sampling, and the efficiency decreases even further by increasing the DEFF. The efficiency of the variance estimates, however, is higher in the two-stage stratified sample in the case of non-zero intradistrict correlations.

$$\lambda = \frac{|\theta - E(\theta)|}{\sqrt{Var(\theta)}}$$

The general conclusion is that it is extremely dangerous to try to assess the prevalence of undernourishment by fitting a log-normal distribution to estimates of mean and variance derived from a two-stage stratified sampling design. In real instances, the situation is probably even worse than that depicted here because the empirical distributions may differ dramatically from the log-normal case. In some preliminary experiments (whose results are not reported here for brevity), we have seen that estimates of the skewness and the kurtosis are also biased and highly inefficient if based on two-stage stratified samples.

TABLE 1. SUMMARY OF RESULTS: MEAN AND VARIANCE						
	Population A[a]		Population B		Population C	
	Relative bias of ST2S [b] (I)	Relative effi ciency of ST2S vs. RS	Relativebias of ST2S (I)	Relative effi ciency of ST2S vs. RS	Relativebias of ST2S (I)	Relative effi ciency of ST2S vs. RS
Mean	0.2550	0.3061	0.5444	0.2962	0.5356	0.2968
Variance	0.1753	0.4550	0.4589	0.4725	0.4549	0.4964

[a] Population A refers to zero intra-district correlation, in population B the intra-district correlation between two urban geographic units (GU) is 0.3, that between an urban GU and any other GU is 0.2 and that between two non-urban GU's is 0.1; in population C the intra-district correlation between two urban GU's is 0.1, that between an urban GU and any other GU is 0.2 and that between two non-urban GU's is 0.3.
[b] ST2S is two-stage stratifi ed sampling; RS is random sampling.

TABLE 2. SUMMARY OF RESULTS: DESIGN EFFECT				
	Mean		Variance	
	Relative bias of ST2S (I)	Relative efficiency of ST2S vs. RS	Relative bias of efficiency ST2S (I)	Relative of ST2S vs. RS
DEFF=0.8[a]	0.5349	0.34	0.4375	0.56
DEFF = 1.0	0.5471	0.30	0.4564	0.49
DEFF = 1.2	0.5976	0.28	0.4675	0.45
DEFF = 1.4	0.4537	0.27	0.4538	0.46
DEFF = 1.6	0.4719	0.26	0.4245	0.38
DEFF = 1.8	0.5760	0.25	0.4814	0.35
DEFF = 2.0	0.5798	0.24	0.4700	0.33

[a] DEFF is the design effect defined as the ratio between the variance within segments and the variance between segments.

In a future paper, we will analyse this point further by assuming skewed-normally distributed populations and populations obeying the four-parameter Pearson family. We will study the quality of estimates of the shape parameters (skewness and kurtosis) and of the lower percentiles based on two-stage stratified samples and also possible solutions to improve the precision of these estimates.

Acknowledgements

The author is indebted to J. Mernies, L. Naiken, P. Narain and R. Sibrian of ESSA-FAO, for the many useful comments and suggestions received on previous versions of this paper.

References

Cochrane, W.G. 1963. *Sampling techniques.* New York, John Wiley.