Teacher's Corner

Sampling from a Skewed Population: The Sampling Design of the 2011 Survey of Enterprises in the Philippines

Erniel B. Barrios
School of Statistics
University of the Philippines Diliman

The sampling design for the 2011 Survey of Enterprises in the Philippines is used to illustrate sampling from a skewed population that is characterized by large population variance. The modified Lavallee-Hidiroglu stratification algorithm is used to simultaneously determine the stratum boundaries and the optimum sample size that will minimize the coefficient of variation. The principles behind "threshold" sampling is then used in excluding segment of the population containing "minimal" information while the segment containing the "most" information is considered as certainty ratum. An estimation procedure is also presented.

Keywords: skewed population, Lavallee-Hidiroglu algorithm, model-assisted estimation

1. Introduction

Business surveys are very challenging for two reasons: (1) the organizational structures of businesses are too complicated, thus it is often difficult to locate where the data is; and (2) the distribution of some target variables like revenue is not normal but rather severely skewed.

The problem of enumerating a firm with complicated organizational structure is addressed by considering an enterprise instead of establishments. Most firms adopt aggregated financial reporting system and hence, by enumerating only the head office, all branches, subsidiaries, and other affiliates of the firm can possibly be accounted properly. This will minimize the bias that could be easily accrued due to coverage problem of establishment-based surveys. This will also minimize the possible item response error for certain indicators that are not necessarily available at the establishment level but is compiled only in the head office.

Skewness of the target variable is addressed by "cut-off" or "threshold" sampling and the use of modified Lavallee-Hidiroglu stratification algorithm. "Cut-off" sampling will ensure that all enterprises contributing "significantly" to the total of the target variable are part of the "must take" stratum and subsequently

considered a certainty stratum. On the other hand, the very small companies that contribute "insignificant" amount to the total of the target variable are classified under the "take none" stratum and no sample will be drawn from this group. While ignoring the "take none" stratum can possibly induce bias, it reduces the burden of enumeration since only few units will be enumerated. The rest of the population are stratified using the Lavallee-Hidiroglu stratification algorithm that aims to minimize both the coefficient of variation and sample size by choosing the stratum boundaries in an iterative process.

2. The Frame

The survey uses a stratified single-stage design, directly enumerating the business enterprises treated as the sampling units. The 2011 Updated List of Establishments (ULE) was used in the aggregation of enterprises. All establishments sharing the first nine digits of Tax Identification Number (TIN) are aggregated into an Enterprise. All establishments without informative TIN or no TIN from ULE are treated as individual enterprises. The enterprise is assigned to a PSIC Section and Region based on the dominance of the total employment (TE) of the establishments under the enterprise. Enterprises whose establishments are classified in more than one PSIC Section regardless of location are considered "Complex" while those enterprises whose establishments fall into one PSIC Section only are considered "Simple."

There are 114,539 enterprises identified from the 2011 ULE (TE of 4,704,451). Of this, 8,784 are complex enterprises accounting for 7.67% of all enterprises (TE of 970,791 accounting for 20.64% of all TE).

3. Sampling Design

The survey follows the framework of "Cut-Off" sampling, see for example, Benedetti et al., (2010). Business enterprises are characterized by skewed distribution where there are too many "small" enterprises contributing very small amount to the total of the target variables. On the other hand, there are very few "large" enterprises contributing large amount to the total of the target variables.

The 18 PSIC Sections are considered as the primary strata or the domains of the survey. Most complex enterprises are also very large in terms of total employment thus, it is often recommended to take all complex enterprises into the survey (Choudhry, 2012), but in case the number is very large, some threshold maybe considered in deciding how many will be included.

To decide on the "must take" stratum, i.e., "large" complex enterprises, the lower quartile for each PSIC section are considered. All complex enterprises whose TE exceeds the lower quartile in the corresponding PSIC section are included in the "must take" stratum except for Sections A, B, D, and E where all complex

enterprises were considered "must take." There are a total of 6,766 enterprises in the "must take" stratum accounting for 77.03% of all complex enterprises or 5.91% of all enterprises.

There are 49,896 (43.62%) "small" enterprises contributing at most 5% in TE. Since their contribution is very small, they are considered as burden in survey operations, thus, generally, they are ignored (Choudhry, 2012).

The "middle" group is further stratified into three strata following Lavallee and Hidiroglou (1988). The stratification method jointly minimizes the Total CV and the sample size in the choice of stratum boundaries and in the determination of sampling rate per stratum. Further accounting for the possible discrepancy between the stratification variables (TE) and the target variable, e.g., revenue, Rivest (2002) modified the Lavallee-Hidiroglou Algorithm (LH Algorithm), see (Horgan, 2006) for a thorough review of methods of stratifying skewed populations. We used the modified LH Algorithm to identify stratum boundaries of the three strata for the "middle" group. Stratum 1 and Stratum 2 are smaller than Stratum 3, and are called the "take some" strata. Stratum 3, comprising of "bigger" enterprises in terms of TE, is also called "take all" stratum. The stratum boundaries vary across the primary strata (PSIC Sections).

The general characteristics of the different strata and the corresponding sampling strategies are summarized in Table 1.

Table 1 Characteristics of the Strata and Sampling Strategies

Stratum	Characteristics	Sampling Strategy
0	Very small enterprises whose aggregate TE contributes at most 5% of TE of all enterprises.	Take None
1	Smallest stratum formed by LH Algorithm	Take Some
2	Medium stratum formed by LH Algorithm	Take Some
3	Largest stratum formed by LH Algorithm	Take All
4	Complex enterprises whose TE exceeds the lower quartile of the Section except A, B, D, and E where all complex enterprises are included	Must Take

In stratum 1 and 2, the enterprises are sorted by Region and Province. Systematic selection of sample enterprises was made to induce implicit stratification among the Regions and Provinces. Thus, while the target domains are the PSIC Sections, the implicitly stratified samples will also facilitate the generation of estimates by Region. This will also ensure that samples are distributed across the Provinces.

The sample size for each production industries in the National Accounts was then evaluated. For industries with very small sample size, the sample was augmented with proportional numbers from Stratums 1 and 2.

4. Sample Size

The survey aims to generate estimates of total value-added, non-financial assets, income/revenue, and gross capital formation for institutional sector accounts of the Philippine System of National Accounts. The frame (ULE), where the establishments are aggregated into their mother enterprises contains TE as the only useful auxiliary information. TE possesses much smaller variation than any of the target variables in the survey. Thus, using TE as the auxiliary variable, the survey aims to achieve a very low coefficient of variation (CV) of the estimates at 4%. Assuming that the target variable exhibits higher variability than TE, it is expected that it would still yield lower CV since a very low CV was considered for TE. This is also the target CV in some countries like Canada who uses TE as auxiliary variable in survey of establishments.

The sample size was determined from the modified LH Algorithm (Benedetti, et al., 2010). The target CV (CV_0) is defines as

$$CV_{specified} = \left(1 + \frac{Y_{must\ take\ complement}}{Y_{must\ take}} \right) CV_0.$$

Let $a_h = \frac{(w_h n_h)^p}{\sum_{h=1}^{l-1} (w_h n_h)^p}$ where $-\infty<p<\infty$, the weighted (w_h) power proportion of sample allocation to the different strata. Stratum l (Stratum 3 in our case) is the "take all" strata since many enterprises in this stratum have very large TE values. The power p is taken as 1, this is known to be fairly robust to further severity of skewness of the distribution of TE.

Then the sample size is computed from

$$n = Nw_l + \frac{N\left[\sum_{h=1}^{l-1} \frac{(w_h\sigma_h)^2}{(w_h\mu_h)^p}\right]\left[\sum_{h=1}^{l-1}(w_h\mu_h)^p\right]}{NCV_{specified}^2\mu^2 + \sum_{h=1}^{l-1}(w_h\sigma_h)^2}$$

The sample size is then allocated to the different strata (1 and 2) using a_h. Note that the sample size is augmented with 10% additional enterprises to cover inaccessible, non-existing, and other non-responding enterprises. The numbers are proportionally distributed to Stratum 1 and 2.

The resulting sample sizes and CV of Stratums 1, 2, and 3 are given in Table 2. Note that the target CV of 4% is achieved in all primary strata (PSIC Sections).

The total sample size from the LH Algorithm is 2,027 or a total of 8,793 including the must take stratum. This will be augmented with 10% for a total of

9,673 enterprises to account for the inaccessible units. The distribution of "must take" enterprises (larger than lower quartile of TE) is given in Table 3.

Table 2 Sample Size and Other Characteristics of Primary Strata

PSIC Section	Sample Size (Stratum 1,2 ,3)	CV of TE		PSIC Section	Sample Size (Stratum 1,2 ,3)	CV of TE
A	79	0.0397		K	121	0.0401
B	13	0.0423		L	124	0.0400
C	215	0.0400		M	117	0.0399
D	26	0.0404		N	88	0.0399
E	58	0.0404		P	191	0.0400
F	78	0.0398		Q	118	0.0400
G	263	0.0400		R	68	0.0398
H	118	0.0400		S	98	0.0400
I	159	0.0400		Total	2027	
J	93	0.0398				

Table 3 Distribution of Must Take by PSIC Sections

PSIC Section	N		PSIC Section	N
A-Agriculture, Forestry, and Fishing	145		K-Financial and Insurance Activities	287
B-Mining and Quarrying	22		L-Real Estate Activities	92
C-Manufacturing	1,481		M-Professional, Scientific and Technical Services	235
D-Electricity, Gas, Steam, and Air Conditioning Supply	20		N-Administrative and Support Services	206
E-Water Supply; Sewerage, Waste Management and Remediation	20		P-Education	135
F-Construction	87		Q-Human Health and Social Work Activities	223
G-Wholesale and Retail Trade	2,058		R-Arts, Entertainment, and Recreation	86
H-Repair of Motor Vehicles and Transportation and Storage	106		S-Other Service Activities	379
I-Accommodation and Food Services Activities	950		TOTAL	6,766
J-Information and Communication	234			

The total sample size distributed by Primary Stratification (PSIC Sections) is given in Table 4 and by Region in Table 5.

The total sample size of 9,673 enterprises comprises 6,766 complex enterprises (Stratum 4), 936 from "Take All" (Stratum 3), and 1,971 from "Take Some" (Stratums 1 and 2).

Table 4 Distribution of Samples by PSIC Section

PSIC Section	Sample Size	Percentage
A-Agriculture, Forestry, and Fishing	257	2.66
B-Mining and Quarrying	39	0.4
C-Manufacturing	1786	18.46
D-Electricity, Gas, Steam, and Air Conditioning Supply	59	0.61
E-Water Supply; Sewerage, Waste Management and Remediation	96	0.99
F-Construction	192	1.98
G-Wholesale and Retail Trade	2463	25.46
H-Repair of Motor Vehicles and Transportation and Storage	269	2.78
I-Accommodation and Food Services Activities	1185	12.25
J-Information and Communication	360	3.72
K-Financial and Insurance Activities	458	4.73
L-Real Estate Activities	269	2.78
M-Professional, Scientific and Technical Services	394	4.07
N-Administrative and Support Services	331	3.42
P-Education	421	4.35
Q-Human Health and Social Work Activities	392	4.05
R-Arts, Entertainment, and Recreation	178	1.84
S-Other Service Activities	524	5.42
Total	9673	99.97

Table 5 Distribution of Samples by Region

Region	Sample Size	Percentage		Region	Sample Size	Percentage
1	311	3.22		10	442	4.57
2	171	1.77		11	697	7.21
3	621	6.42		12	399	4.12
4	888	9.18		13	3625	37.48
5	274	2.83		14	103	1.06
6	537	5.55		15	42	0.43
7	751	7.76		16	167	1.73
8	150	1.55		17	196	2.03
9	299	3.09		Total	9673	100

5. Estimation Procedure

The population is divided into three groups: U_C=Take All (Stratum 3 and 4); U_S=Take Some (Stratum 1 and 2); and U_E= "Take None" (Stratum 0). Following (Benedetti et al., 2010), let Y=target variable, X=auxiliary variable. The estimate of the total Y is the sum of the completely enumerated enterprises, estimates from the sampled enterprises, and an estimate (imputation) for the non-sampled enterprises. The estimate of the total $\hat{t}_Y$ is

$$\hat{t}_Y = t_C + \hat{t}_S + \hat{t}_E$$

where

t_C= total of certainty units (must take+take all)

$\hat{t}_S = \sum_{k \in S} \frac{Y_K}{\pi_K}$ = estimator in take some (Stratum 1 and 2) strata

$\hat{t}_E$ = is the estimator of take none stratum (Stratum 0)

Let $\delta = \frac{\sum_{k \in U_E} Y_k}{\sum_{k \in U_C} Y_k + \sum_{k \in U_S} Y_k}$ and $\tilde{\delta} = \frac{\sum_{k \in U_E} X_k}{\sum_{k \in U_C} X_k + \sum_{k \in U_S} X_k}$

Then $\hat{t}_Y = \left(1 + \tilde{\delta}\right)\left[\sum_{k \in U_C} Y_k + \sum_{k \in U_S} w_k Y_k\right]$ where w_k=sampling weight (also included in the list of sample enterprises).

$\hat{t}_Y$ = is model-assisted estimator and is biased. The MSE is computed instead of variance given as follows:

$$MSE\left(\hat{t}_Y\right) = \left(1 + \tilde{\delta}\right)^2 V\left(\hat{t}_S\right) + Bias^2\left(\hat{t}_Y\right)$$

where $Bias^2\left(\hat{t}_Y\right) = \left(\tilde{\delta} - \delta\right)\left(\hat{t}_C + \hat{t}_S\right)$

δ is computed from the census of establishments.

Alternatively, the bootstrap method will also be used to address the bias in the estimator and the complexity of the MSE estimator, see (Hidiroglou and Srinath, 1981), (Canty and Davison, 1999), and (Singh et al., 2001).

The non-responding enterprises can be imputed via weighting adjustment (Kalton and Cervantes, 2003).

References

BENEDETTI, R, M. BEE and G. ESPA, 2010, A framework for cut-off sampling in business survey design, Journal of Official Statistics, Vol. 26(4): 651-671.

CANTY, A. and A. DAVISON, 1999, Resampling-based variance estimation for labour force surveys, Journal of the Royal Statistical Society (D), 48(3): 379-391.

CHOUDHRY, G., 2012, Report on the development of a sample design for the survey of enterprises in the Philippines, Technical Consultancy Report Submitted to the National Statistics Office.

HIDIROGLOU, M. and K. SRINATH, 1981, Some estimators of a population total from Simple Random Samples containing large units, Journal of the American Statistical Association, 76(375): 690-695.

HORGAN, J., 2006, Stratification of skewed populations: A review, International Statistical Review, 74(1):67-76.

KALTON, G. and I. CERVANTES, 2003, Weighting methods, Journal of Official Statistics, 19(2): 81-97.

LAVALLEE, P. and M. HIDIROGLOU, 1988, On the stratification of skewed populations, Survey Methodology, 14:33-43.

RIVEST, L., 2002, A generalization of the Lavallee and Hidiroglou algorithm for stratification in business surveys, Survey Methodology, 28(2):191-198.

SINGH, M., M. HIDIROGLOU, J. GAMBINO, and M. KOVACEVIC, 2001, Estimation methods and related systems at Statistics Canada, International Statistical Review, 69(3): 461-485.

Acknowledgement

The author expresses his gratitude to Dir. Estela De Guzman, Industry and Trade Statistics Department, National Statistics Office, for allowing him to include part of his consultancy report in this paper. Joseph Ryan G. Lansangan is acknowledged for programming assistance.