Estimation Under Purposive Sampling with Auxiliary Variable

John Erwin Bañez
University of the Philippines Diliman

A PPS purposive selection and estimation procedure is explored using purposive sampling proposed by Guarte and Barrios (2006). Instead of SRS, PPS with auxiliary variable was used in the selection. Results were compared to estimates from ordinary SRS and SRS purposive selection with standard error and coefficient of variation of estimates as basis. PPS purposive has comparable results to SRS purposive and both are better compared to ordinary SRS.

Keywords: Purposive sampling, bootstrap, auxiliary variable

1. Introduction

Purposive sampling is widely used in the social sciences. It has varied implementation procedures, ranging from convenience sampling, where the selection of units is entirely up to the enumerator, up to purposive selection of segments where random sampling will be implemented.

This paper follows the work of Guarte and Barrios (2006) on purposive sampling. Purposive sampling is defined as,

> "randomly selecting units without replacement from the particular section of the population believed to yield samples that will give best estimate of the population parameter of interest" (Guarte and Barrios, 2006).

Guarte and Barrios (2006) proposed a purposive sampling (selection and estimation) procedure. They used simple random sampling (SRS) only on the section of population believed to provide more precise information about the population mean. Their procedure was tested on datasets with different parameters and distribution. They used bootstrap procedure and trimming of less informative units in the population to estimate the population mean. They concluded that their proposed procedure provides biased but more precise estimated mean.

This paper implements the Guarte and Barrios (2006) method on a normally distributed data with varying mean and variances. We explore the effect of auxiliary variable to the estimation procedure. Essentially, this means using probability proportional to size (PPS) where the size measure is the auxiliary variable – a variable with a positive correlation with the variable of interest. A practical application may be when an estimation of children under 15 years old is needed. The researcher may deduce from experience and from past studies that the number of children under 15 years of age is positively correlated to the population size. Similarly estimation of poverty incidence may also benefit from PPS since poverty incidence may be positively correlated to population size.

2. Simulation

For comparability, similar simulation settings presented by Guarte and Barrios (2006) were used in generating population and in estimation of $\bar{Y}$. A slight modification is, in PPS purposive, selection of samples was influenced by an auxiliary variable X whose correlation with Y was made to vary from 0.3 (low) to 0.8 (high).

2.1 Simulated population

We create datasets of size N=500 with the following characteristics:

Dataset 1: Homogeneous and low correlation (low hom)
Normally distributed with mean ≈ 20 and variance ≈ 1
Correlation between Y and X is low (r= 0.3)

Dataset 2: Homogeneous and high correlation (high hom)
Normally distributed with mean ≈ 20 and variance ≈ 1
Correlation between Y and X is high (r= 0.8)

Dataset 3: Heterogeneous and low correlation (low het)
Normally distributed with mean≈ 16 and variance≈ 2500
Correlation between Y and X is low (r= 0.3)

Dataset 4: Heterogeneous and high correlation (high het)
Normally distributed with mean≈ 16 and variance≈2500
Correlation between Y and X is high (r= 0.8)

2.2 Selection procedure

We modify the selection procedure proposed by Guarte and Barrios (2006). This time, probability proportional to size (PPS) of auxiliary variable is used instead of simple random sampling (SRS). The selection procedure is stated below:

Purposive sampling is defined as randomly selecting units with probability of selection proportional to size (PPS) of auxiliary variable without replacement

from the particular section of the population believed to yield samples that will give the best estimate of the population parameter of interest. The goal is to estimate the population mean.

Using PPS, randomly draw without replacement from the central location of the distribution. Other segments (less informative) of the population were assigned zero probability of inclusion. The auxiliary variable serves as the size measure.

3. Empirical Implementation

The target population is trimmed of the tail values, that is, values less than $\left(\theta-\sqrt{\frac{\sigma^2}{2}}\right)$ and values greater than $\left(\theta+\sqrt{\frac{\sigma^2}{2}}\right)$ were deleted. Note that trimming will be done only to heterogeneous populations (i.e. Dataset 3 and 4).

3.1 Estimation procedure

Let S_i be a bootstrap replication from the original samle S. Let

$$\pi_i=\begin{cases} n\dfrac{x_i}{\sum_{i=1}^{N} x_i'} & \theta-\sqrt{\dfrac{\sigma^2}{2}}<X_i<\theta+\sqrt{\dfrac{\sigma^2}{2}} \\ 0, & \textit{otherwise} \end{cases} \tag{1}$$

Then, for the i[th] bootstrap replicate

$$\bar{\theta}_{S_i}=\sum_{j\in S_i}\frac{Y_j}{\pi_j} \tag{2}$$

The bootstrap estimates in then

$$\bar{\theta}^*=\sum_{i-1}^{k}\bar{\theta}_{s_i} \tag{3}$$

and the standard error is

$$s.e.\left(\bar{\theta}^*\right)=\left[\frac{1}{k-1}\sum_{j=1}^{k}\left(\bar{\theta}_{S_i}-\bar{\theta}^*\right)^2\right]^{1/2} \tag{4}$$

The presence of bias in sample selection will be addressed. A bias-corrected method for computing 95% confidence interval (CI) as discussed by Guarte and Barrios (2006) was computed. This method was chosen since it has the weakest parametric assumptions.

4. Plan of Analysis

Estimated mean, standard error (s.e.) and coefficient of variation (c.v.) from SRS, PPS purposive and SRS purposive (as in Guarte and Barrios, 2006) will be compared. A bias-corrected 95% confidence interval for SRS purposive and PPS purposive will also be compared.

Table 1 Summary of Simulated Data

Data	Mean	Variance	cut-off1	cut-off2
low hom (pearson's r = .3; variance = 1)	19.94	.097		
high hom (pearson's r = .8; variance = 1)	19.94	0.94		
low het (pearson's r = .3; variance = 2500)	16.78	2422.23	-18.02	51.58
high het (pearson's r = .8; variance = 2500)	16.78	2351.40	-17.51	51.06

5. Results and Discussion

The estimates from PPS Purposive have generally smaller c.v. and s.e. This is true both for homogeneous and heterogeneous populations. The difference in c.v. of PPS purposive estimates of homogeneous and heterogeneous populations is very small, compared to the SRS estimates.

SRS purposive yielded estimates with small s.e. similar to PPS purposive. The c.v. of PPS purposive were found to be similar to that of SRS purposive for these populations. The bias and s.e. is not significantly affected by increasing sample size. The estimates of the PPS purposive is efficient and consistent both in the homogeneous and heterogeneous populations.

Thus, while the presence of an auxiliary variable may help, this is not necessarily an improvement over SRS purposive sample selection. The ease of using SRS compared to PPS makes SRS purposive a more viable method.

Since no auxiliary variable is used in SRS we disregard the pearson r. Thus, we implement the estimation procedure to heterogeneous population only (dataset 4).

Similar to the results of Guarte and Barrios, the bias is affected by the resample size. The shortest interval was given by the largest resample size. This is true for SRS Purposive and PPS purposive estimates.

Table 2 **Comparison of Ordinary SRS and PPS Purposive Sampling Estimates when the Population is Homogeneous: N(20,1) with r ≈ 0.3 and N = 500**

	Ordinary SRS			**PPS Purposive**		
Sample size n	mean	se	cv	mean	se	cv
15	19.946	0.205	3.985	19.937	0.026	0.001
30	19.806	0.176	4.861	19.951	0.018	0.001
50	19.929	0.149	5.280	19.919	0.014	0.001
100	19.977	0.109	5.445	19.946	0.010	0.000

Table 3 **Comparison of Ordinary SRS and PPS Purposive Sampling Estimates when the Population is Homogeneous: N(20,1) with r ≈ 0.8 and N = 500**

	Ordinary SRS			**PPS Purposive**		
Sample size n	mean	se	cv	mean	se	cv
15	19.946	0.178	3.455	19.932	0.025	0.001
30	19.792	0.168	4.663	19.936	0.018	0.001
50	19.864	0.128	5.562	19.916	0.014	0.001
100	19.844	0.104	5.244	19.956	0.010	0.000

Table 4 **Comparison of Ordinary SRS and PPS Purposive Sampling Estimates when the Population is Heterogeneous: N(16,2500) with r ≈ 0.3 and N = 500**

	Ordinary SRS			**PPS Purposive**		
Sample size n	mean	se	cv	mean	se	cv
15	17.290	10.261	229.851	15.293	0.0201%	32.520
30	10.294	8.788	467.601	15.328	0.0288%	22.682
50	16.462	7.441	319.637	14.916	0.0378%	17.750
100	18.836	5.439	288.750	15.044	0.0625%	10.641

Table 5 Comparison of Ordinary SRS and PPS Purposive Sampling Estimates when the Population is Heterogeneous: N(16,2500) with pearson r ≈ 0.8 and N = 500

	Ordinary SRS			PPS Purposive		
Sample size n	**mean**	**se**	**cv**	**mean**	**se**	**cv**
15	17.290	8.897	199.281	16.665	0.0208%	28.856
30	9.587	8.424	481.268	16.336	0.0030%	20.242
50	13.206	6.408	343.135	16.426	0.0393%	15.511
100	12.209	5.203	426.126	16.310	0.0649%	9.454

Table 6 Estimates from Purposive SRS when the Population is Heterogeneous: N(16,2500) and N = 500

	SRS trimmed		
Sample size n	**mean**	**se**	**cv**
15	16.557	0.0666%	28.843
30	16.492	0.1024%	18.824
50	15.995	0.1143%	17.391
100	16.340	0.1960%	9.928

Note: resample size = n

Table 7 Bias-corrected 95% CI of Ordinary SRS Estimates

Sample size	sample-mean	bootstrap mean	bias	p_lb	p_ub	width
15	15.293	15.293	0.000	6.095	25.803	19.708
30	15.328	15.329	0.000	8.903	22.934	14.030
50	14.916	14.916	0.000	9.814	20.198	10.384
100	15.044	15.044	0.000	11.703	18.082	6.379

Note: resample size = n

Table 8 Bias-corrected 95% CI of Purposive PPS Estimates when the Population is Heterogeneous: N(16,2500) with r ≈ 0.3 and N = 500

Sample size	sample-mean	bootstrap mean	bias	p_lb	p_ub	width
15	16.665	16.665	0.000	7.903	26.928	19.026
30	16.336	16.336	0.000	10.115	22.458	12.344
50	16.426	16.426	0.000	11.030	20.239	20.945
100	16.310	16.310	0.000	13.325	19.309	5.984

Note: resample size = n

Table 9 Bias-corrected 95% CI of Purposive SRS Estimates when the Population is Heterogeneous: N(16,2500) with pearson r ≈ 0.8 and N = 500

Sample size	sample-mean	bootstrap mean	bias	p_lb	p_ub	width
15	16.097	16.099	0.234%	5.280	25.211	19.931
30	16.321	16.321	0.034%	10.325	23.690	13.365
50	15.946	16.947	0.069%	9.242	19.003	9.760
100	16.426	16.426	0.029%	13.811	19.350	5.538

Note: resample size = n

6. Conclusion

Results support the proposed purposive method by Guarte and Barrios. Purposive PPS with auxiliary variable estimation provides similar results to purposive SRS estimation. This highlights the strength of SRS which is its ease of implementation. In cases where an auxiliary variable is worth using or should be used (as in stratification), the purposive estimation shown here is useful.

REFERENCES

GUARTE, J. and E. BARRIOS, 2006, Estimation under purposive sampling, *Communications in Statistics – Simulation and Computation*, 35(2):277-284.

SAS FAQ: How can I bootstrap estimates in SAS?. Available at: http://www.ats.ucla.edu/stat/sas/faq/bootstrap.htm

Sample 25008: Generate data from a multivariate normal distribution.Available at:http://support.sas.com/kb/25/008.html

%MVN macro: Generating multivariate normal data.Available at: http://www.math.clemson.edu/~calvinw/MthSc405/sas/mvn.sas