ASYMPTOTIC DISTRIBUTION OF ESTIMATORS FROM UNEQUAL PROBABILITY SAMPLING

Mohammad A. Chaudhary, Pranab K. Sen, University of North Carolina at Chapel Hill
Mohammad A. Chaudhary, Department of Biostatistics, UNC-CH, Chapel Hill, NC 27599-7400

KEY WORDS: Large Sample Theory, Finite Population, Successive Sampling

If sampling is done with replacement, the sample estimates derived from large surveys can be assumed to approximate normal distribution providing a valid base for statistical inference. With the use of unequal probability without replacement sampling schemes at the first stage, the estimated primary unit totals $\hat{Y}_i$ will no more be independent. Therefore in single as well as multiple stage sample designs involving the use of unequal probability without replacement sampling schemes, the general linear estimator of population total or mean may not follow the normal distribution. Consequently, the estimation of confidence intervals and tests of hypotheses may not be based on the normal approximation assumptions.

The work presented in this manuscript builds on Rosen's (1972) results and pertains to establishing in a more direct way that for large scale sample surveys under some general conditions on the size of units, the general linear estimator of population total from the single stage cluster sample designs using unequal probability without replacement sampling, approximates the normal distribution.

1. Introduction

If the number of psu's N is small, the sampling distribution of the estimator may be evaluated by considering all possible samples of a given size n. For large N (and n) the process becomes prohibitively laborious. In this case, Madow (1949) introduced the permutational central limit theorem for finite population sampling. Rosen (1972) studied the asymptotic nomality for successive sampling without replacement through the coupon collector's problem. Building upon Rosen's work, Sen (1979, 1980) studied the asymptotic distribution theory of estimates of finite population total in single and multistage sampling with varying probabilities without replacement using martingale approach and in relation to the extended coupon collectors problem. However for unequal probability case, a lot of work towards asymptotic normality remains to be done (Sen, 1988).

The next section provides the theoretical details of the asymptotic normality results.

2. Asymptotic Distribution

Successive sampling without replacement refers to drawing units one after the other without replacement from a population of size N. At each draw the probability that the unit i is drawn is proportional to the single draw probability p_i if it remains in the population and zero otherwise. Let $\hat{Y}$ denote the estimator of population total Y under successive sampling without replacement and $\hat{Y}^*$ the corresponding estimator from successive sampling with replacement sampling discarding previously selected units. For the later scheme of sampling, let m_n denote the random sample size or waiting time in Rosen's terminology to obtain n distinct units. Rosen (1972) showed that $\hat{Y}$ and $\hat{Y}^*$ are equivalent in distribution i.e.,

$$\hat{Y} \approx \hat{Y}^* \tag{2.1}$$

Rosen also provided the asymptotic approximations for the inclusion probabilities π_i, its variance and covariance when n and N are large and under some additional general conditions on the size of units. The approximation to the inclusion probabilities entails replacing successive sampling without replacement by the one with replacement by adjusting the sample size so that variance remains the same asymptotically. With this adjustment the resulting approximate estimator $\hat{Y}''$ becomes a linear combination of independent random variables.

The general strategy adopted in this paper to obtain the asymptotic normality results for the linear estimator $\hat{Y}$ involves the following three steps: (i) Show that $\hat{Y}''$ and $\hat{Y}^*$ are convergent equivalent, i.e.,

$$\frac{1}{(n)^{1/2}}\left(\hat{Y}'' - \hat{Y}^*\right) \xrightarrow{p} 0 \tag{2.2}$$

(ii) Show that ratio of the variance of the general linear estimator $\hat{Y}'$ and the variance of the estimator $\hat{Y}''$ goes to 1 i.e.,

$$V(\hat{Y})/V(\hat{Y}') \to 1 \qquad (2.3)$$

(iii) Show that the standardized form of $\hat{Y}'$ satisfies the Liapounov condition and therefore approximates the standard normal distribution so that,

$$\hat{Y}' \sim N\left[E(\hat{Y}'), V(\hat{Y}')\right] \to 1 \qquad (2.4)$$

The result (i) may be proved easily either by showing that Anscombe condition [Anscombe(1952), Sen & Singer (1993), Example, 8.3.1] is satisfied or by using martingale approach. Section (2.1) deals with (ii) and section (2.2) with (iii).

2.1 Convergence of Variance

$$\text{Let} \quad \hat{Y}' = \hat{Y} - Y = \sum_{i=1}^{N} \frac{w_i' Y_i}{\pi_i} - Y$$

$$= \sum_{i=1}^{N} \left(\frac{w_i'}{\pi_i} - 1 \right) Y_i , \qquad (2.5)$$

where $\hat{Y}$ is the Horvitz-Thompson estimator of the population total Y and Y_i is the i^{th} psu total. Also $w_i' = 1$ if i^{th} unit is selected in the sample of size n and zero otherwise and

$$\begin{aligned} P(w_i' = 1) &= \pi_i \quad \text{and} \\ P(w_i' = 0) &= 1 - \pi_i \end{aligned} \qquad (2.6)$$

$E(\hat{Y}') = 0$ and variance is given by the well known Yates-Grundy variance formula:

$$V(\hat{Y}') = \sum_{i<j}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_i}{\pi_i} \right)^2 \qquad (2.7)$$

Hartley and Rao (1962) and Rao (1963) provided the following approximation for $V(\hat{Y}')$ correct to $O(N)$:

$$V(\hat{Y}') = \sum_{i=1}^{N} \frac{1}{\pi_i} \left(1 - \frac{n-1}{n} \pi_i \right) \left(Y_i - \frac{\pi_i Y}{n} \right)^2 \qquad (2.8)$$

This approximation is good for (i) Narain's Method [Narain (1951), Yates & Grundy, (1953)], (ii) PPS Systematic Random Sampling (Goodman & Kish, 1950) and (iii) Yates-Grundy Rejective Sampling [Yates and Grundy (1953)]. For each of these three procedures $\pi_i = np_i$. For a discussion of these and other unequal probability sampling procedures, see Brewer & Hanif (1983).

Given that a combination (N,n) and the corresponding $(p_1, p_2 \ldots, p_N)$ can be regarded as an element of the sequence of sampling situations $\{(\mathrm{p_k}, \pi_k)\}_{k=1}^{\infty}$ where $\mathrm{p_k} = (p_{k1}, p_{k2} \ldots, p_{kN})$, $\pi_k = (\pi_{k1}, \pi_{k2} \ldots, \pi_{kN})$ and $\sum_{i=1}^{N} p_{ki} = 1$. Consider the following two conditions:

$$\lim_{k \to \infty} N_k = \infty \qquad (2.9)$$

$$\limsup_{k \to \infty} \frac{\max_i p_{ki}}{\min_i p_{ki}} < \infty \qquad (2.10)$$

where N_k is the population size, p_{ki} is the single draw probability of the i^{th} unit and π_{ki} is the inclusion probability of the i^{th} unit in a sample of size n_k corresponding to the k^{th} sampling situation. We shall drop the subscript k for the simplicity of the presentation unless essential.

If (2.9) & (2.10) are satisfied, then the approximation for π_i is given by,

$$\pi_i'' = 1 - e^{-p_i t(n)} , \qquad (2.11)$$

where the function $t(x)$ may be termed as the revised sample size and is implicitly defined by the relationship.

$$N - x = \sum_{i=1}^{N} e^{-p_i t(x)} \qquad (2.12)$$

Using this approximation, let

$$\hat{Y}'' = \sum_{i=1}^{N} \frac{w_i''}{\pi_i''} (Y_i - \overline{Y})$$

$$= \sum_{i=1}^{N} w_i'' \left(1 - e^{-p_i t(n)}\right)^{-1} (Y_i - \overline{Y}), \qquad (2.13)$$

where $\overline{Y} = Y/N$. Also $w_i'' = 1$ if i^{th} unit is selected in the sample of size n and zero otherwise and are independent for $i = 1,2,\ldots,N$ and

$$P(w_i'' = 1) = \pi_i'' = 1 - e^{-p_i t(n)} \quad \text{and}$$
$$P(w_i'' = 0) = 1 - \pi_i'' = e^{-p_i t(n)}. \tag{2.14}$$

$E(\hat{Y}'') = 0$ and

$$V(\hat{Y}'') = \sum_{i=1}^{N} \frac{1-\pi_i''}{\pi_i''}(Y_i - \overline{Y})^2 \tag{2.15}$$

The estimator $V(\hat{Y}'')$ provides an approximation for $V(\hat{Y})$ which may easily be realized in simple random sampling case where $p_i = 1/N$, $\pi_i = n/N$ so that the equation (2.15) reduces to:

$$V(\hat{Y}'') = \frac{N-n}{n}\sum_{i=1}^{N}(Y_i - \overline{Y})^2,$$

$$= \frac{N-1}{N}\left[N^2(1-f)\frac{S^2}{n}\right] \tag{2.16}$$

where $f = n/N$. The factor $(N-1)/N$ vanishes out when N is large.

With the use of some additional general conditions, it can be shown that $V(\hat{Y}')$ converges to $V(\hat{Y}'')$ in the unequal probability sampling case. Rewriting (2.8) as

$$V(\hat{Y}') = \sum_{i=1}^{N}\frac{1}{\pi_i}\left(1 - \frac{n-1}{n}\pi_i\right)\left(Y_i - \overline{Y} + \overline{Y} - \frac{\pi_i Y}{n}\right)^2$$

$$= \sum_{i=1}^{N}\frac{1}{\pi_i}\left(1 - \frac{n-1}{n}\pi_i\right)\left[(Y_i - \overline{Y})^2 + Y^2\left(\frac{1}{N} - \frac{\pi_i}{n}\right)^2 + 2Y(Y_i - \overline{Y})\left(\frac{1}{N} - \frac{\pi_i}{n}\right)\right] \tag{2.17}$$

and dividing by (2.15),

$$\frac{V(\hat{Y}')}{V(\hat{Y}'')} = \frac{1}{V(\hat{Y}'')}\sum_{i=1}^{N}\frac{1}{\pi_i}\left(1 - \frac{n-1}{n}\pi_i\right)(Y_i - \overline{Y})^2$$

$$+\frac{Y^2}{N^2 V(\hat{Y}'')}\left[\sum_{i=1}^{N}\frac{1}{\pi_i}\left(1 - \frac{n-1}{n}\pi_i\right)(Np_i - 1)^2\right]$$

$$+\frac{2Y}{NV(\hat{Y}'')}\sum_{i=1}^{N}\frac{1}{\pi_i}\left(1 - \frac{n-1}{n}\pi_i\right)(Y_i - \overline{Y})(1 - Np_i) \tag{2.18}$$

Since,

$$\frac{1}{\pi_i}\left(1 - \frac{n-1}{n}\pi_i\right) = \frac{1-\pi_i}{\pi_i} + O(N^{-1}) \tag{2.19}$$

Replacing π_i by π_i'', the first term of (2.18) goes to 1. The second term may be written as:

$$V_2 = \frac{Y^2 W}{N^2 V(\hat{Y}'')}\left[\sum_{i=1}^{N}\frac{W_i}{W}(Np_i - 1)^2\right]$$

$$= \frac{Y^2 W}{N^2 V(\hat{Y}'')}\sum_{i=1}^{N}\frac{W_i}{W}\left[(Np_i - N\overline{p}_w) + (N\overline{p}_w - 1)\right]^2$$

$$= \frac{Y^2 W}{N^2 V(\hat{Y}'')}\left[N^2\overline{p}_w^2\sum_{i=1}^{N}\frac{W_i}{W}\frac{(Np_i - N\overline{p}_w)^2}{N^2\overline{p}_w^2} + (N\overline{p}_w - N\overline{p})^2\right]$$

$$= \frac{Y^2 W}{N^2 V(\hat{Y}'')}\left[N^2\overline{p}_w^2 \mathrm{CV}(Np_i) + (N\overline{p}_w - N\overline{p})^2\right]$$

where $W_i = \frac{1}{\pi_i}\left(1 - \frac{n-1}{n}\pi_i\right)$, $W = \sum_{i=1}^{N} W_i$ so that

$\frac{1}{W}\sum_{i=1}^{N} W_i = 1$, $\overline{p}_w = \sum W_i p_i / W$, $\overline{p} = \sum p_i / N = 1/N$ and $\mathrm{CV}(Np_i)$ is a weighted coefficient of variation of Np_i values with respect to W_i.

Under the restriction that Np_i values would only be varying in close proximity of 1, we assume that $|N\overline{p}_W - N\overline{p}| = o(1)$ and $\mathrm{CV}(Np_i) = o(1)$. It is to note that $p_i = O(N^{-1})$, $\pi_i = O(1)$ and $W_i = O(1)$ so that $W = O(N)$. Given these results, it is easy to verify that $\overline{p}_W = O(N^{-1})$ making the whole expression inside the square bracket $o(1)$. Also since Y and $V(\hat{Y}'')$ are $O(N)$, the factor outside the square brackets is $O(1)$ resulting in the entire second term to be $o(1)$. We may write as:

$$\sum_{i=1}^{N}\frac{W_i}{W}(Np_i-1)^2=o(1) \qquad (2.20)$$

The third term may be written as:

$$V_3=\frac{2YW}{NV(\hat{Y}'')}\left[\sum_{i=1}^{N}\frac{W_i}{W}(Y_i-\overline{Y})(1-Np_i)\right]$$

The expression outside the square bracket is O(1). A reference to the Cauchy-Schawartz inequality immediately shows that the expression inside the square brackets is negligible. According to the inequality:

$$\left[\sum_{i=1}^{N}\frac{W_i}{W}(Y_i-\overline{Y})(1-Np_i)\right]^2 \leq$$

$$\left[\sum_{i=1}^{N}\frac{W_i}{W}(Y_i-\overline{Y})^2\right]\left[\sum_{i=1}^{N}\frac{W_i}{W}(Np_i-1)^2\right] \qquad (2.21)$$

The first expression on the right hand side is O(1). From (2.20) the second expression is $o(1)$ making the entire right hand side of (2.21) to be $o(1)$. Moreover since W_i and therefore W are finite and positive quantities,

$$\left[\sum_{i=1}^{N}\frac{W_i}{W}(Y_i-\overline{Y})(1-Np_i)\right]^2=o(1),$$

$$\Rightarrow \qquad \sum_{i=1}^{N}\frac{W_i}{W}(Y_i-\overline{Y})(1-Np_i)=o(1) \qquad (2.22)$$

We have seen that the first term converges to 1 and the second and third are o(1) concluding that $V(\hat{Y}'')$ approximates $V(\hat{Y}')$ under the conditions that $n\to\infty$ and $N\to\infty$ simultaneously and p_i's vary only in the close proximity of $1/N$.

2.2 Asymptotic Distribution

The estimator $\hat{Y}''$ is the sum of independently but not identically distributed random variables Y_i'' where

$$Y_i''=w_i''(1-e^{-p_i t(n)})^{-1}(Y_i-\overline{Y}) \qquad (2.23)$$

$$E(Y_i'')=Y_i-\overline{Y}=\mu_i'' \qquad (2.24)$$

$$V(\hat{Y}_i'')=\frac{e^{-p_i t(n)}}{1-e^{-p_i t(n)}}(Y_i-\overline{Y})^2=\sigma_i''^2 \qquad (2.25)$$

Also $E(\hat{Y}'')=0$ and $V(\hat{Y}'')$ is given by (2.15).

Let $$\hat{Z}''=\hat{Y}''/\left[V(\hat{Y}'')\right]^{1/2} \qquad (2.26)$$

In order for $\hat{Z}''$ to approximate the standard normal distribution, we need to show that the Liapounov condition is satisfied. (see Sen & Singer (1993), Theorem 3.3.2).

Introducing the subscript k, consider again the infinite sequence of sampling situations $\{(\mathrm{p}_k,\pi_k)\}_{k=1}^{\infty}$ where $\mathrm{p}_k=(p_{k1},p_{k2}\ldots,p_{kN})$ and $\pi_k=(\pi_{k1},\pi_{k2}\ldots,\pi_{kN})$ defined earlier. Assuming that for some $\delta>0$ the moments of order $2+\delta$ exist and have bounded limits as $\lim_{k\to\infty} N_k=\infty$, i.e.,

$$\limsup_{k\to\infty}\ \frac{1}{N_k}\sum_{i=1}^{N_k}\left|y_{ki}-\overline{Y}_k\right|^{2+\delta}<\infty \qquad (2.27)$$

$$\Rightarrow \quad \limsup_{k\to\infty}\ \max_{1\le i\le N_k}\left|y_{ki}-\overline{Y}_k\right|^{2+\delta}=O(N_k)$$

$$\Rightarrow \quad \limsup_{k\to\infty}\ \max_{1\le i\le N_k}\left|y_{ki}-\overline{Y}_k\right|=O(N_k^{(2+\delta)^{-1}})$$

$$\Rightarrow \quad \limsup_{k\to\infty}\ \frac{1}{N_k}\ \max_{1\le i\le N_k}\left(y_{ki}-\overline{Y}_k\right)^2$$

$$=O(N_k^{-\delta/(2+\delta)})=o(1) \qquad (2.28)$$

This proves the Liapounov condition implying that

$$\hat{Z}''\sim \mathrm{N}(0,\ 1) \qquad (2.29)$$

$$\text{or} \qquad \hat{Y}''\sim \mathrm{N}\left(0,\ \sum\sigma_i''^2=V(\hat{Y}'')\right) \qquad (2.29)$$

Note that N_k is the population size, y_{ki} is the i^{th} observation and $\overline{Y}_k$ is the population mean corresponding to the k^{th} sampling situation.

This completes the proof that,

$$\hat{Y}'\sim \mathrm{N}\left(0,\ V(\hat{Y}')\right) \qquad (2.29)$$

where $V(\hat{\bar{Y}})$ is given by (2.7).

Acknowledgments: The first author was supported by a scholarship from the Government of Pakistan and a corresponding study leave from the Punjab University.

3. References

Anscombe, F.J., (1952) "Large Sample Theory of Sequential Estimation", *Proceedings of the Cambridge Philosophical Society*, **48**, 600-607.

Brewer and Hanif, (1983), "Unequal Probability Without Replacement Sampling" Springer Verlag New York Inc.

Goodman, R. and Kish, L., (1950), "Controlled Selection - a Technique in Probability Sampling", *Journal of American Statistical Association*, **45**, 350-372.

Hartley, H.O., and Rao, J.N.K. (1962), "Sampling With Unequal Probabilities and Without Replacement", *Annals of Mathematical Statistics*, **33**, 350-374.

Madow, W.G., (1949), "On the Theory of Systematic Sampling, II", *Annals of Mathematical Statistics*, **20**, 333-354.

Narain, R.D., (1951), "On Sampling Without Replacement with Varying Probabilities", *Journal of Indian Society of Agricultural Statistics*, **3**, 169-175.

Rao, J.N.K., (1963), "On Three Procedures of Unequal Probability Sampling Without Replacement", *Journal of American Statistical Association*, **58**, 202-215.

Rosen, B. (1972), "Asymptotic Theory for Successive Sampling With Varying Probabilities Without Replacement, I and II", *Annals of Mathematical Statistics*, **43**, 373-397; 748-776.

Sen, P.K. (1979), "Invariance Principles of the Coupon Collector's Problem: A Martingale approach", *Annals of Statistics*, Vol. 7, No. 2, 372-380.

Sen, P.K. (1980), "Limit Theorems for an Extended Coupon Collector's Problem and for Successive Subsampling with Varying Probabilities", *Culcutta Statistical Association Bulletin*, Vol. 29, 113-132.

Sen, P.K. (1988), "Asymptotics in Finite Population Sampling", *Hand Book of Statistics 6*, P.R. Krishnaiah and C.R., Rao Eds., Elsevier Science Publishers B.V, 291-331.

Sen, P.K., and Singer, J.M. (1993), *Large Sample Methods in Statistics: An Introduction With Applications*, Chapman & Hall, NY.

Yates, F., and Grundy, P.M. (1953) "Selection Without Replacement From Within Strata With Probability Proportional to Size", *Journal of the Royal Statistical Society, Ser. B*, **15**, 253-261.