On the estimation of omission rate for Indian census count

Kiranmoy Chatterjee*†
Indian Statistical Institute, Kolkata

1 Introduction

Census data is the only primary and complete source of demographic data and SRS(Sample registration System) estimates vital events in India. These two takes an important role in formulation of the socio-economy related policies and planning by Indian government. Inevitably, different types of counting errors exist in census. To estimate the counting error in census population size or total number of any vital events in India two data sources are used generally. CD estimator due to Chandrasekar and Deming(1949, [3]), which is popularly known as Dual System Estimate(DSE) in demography(Alho and Spencer 2005, [2]), is used for estimating omission in census. SRS also uses two sources of data - one is continuous (longitudinal) enumeration of vital events and another is independent retrospective half-yearly surveys, to estimate births and deaths. In this article we will give emphasis on the DSE method which is being criticized since last three decades. As per our knowledge, no constructive discussion has been made yet on the methodological issues in Indian context. Estimating the exact size of population and vital statistics are very much essential for effective policy formulation and implementation. A specialized survey, known as Post Enumeration Survey(PES), conducted within three months of Indian census enumeration and is used to estimate omission rates. Assessment of methodological and operational aspects of Indian coverage error estimation has enough scope to strengthen the evaluation process to have better quality information from census. Bias due to the presence of heterogeneity and/or dependence between capture-recapture probabilities are discussed by several statisticians and demographers(see [3],[13], [4], and [12]). To ensure homogeneity, Chandrasekar and Deming[3] suggested to form poststrata dividing the population according to various cross-sectional age, race, sex and geographical groups. An estimate of coverage error in each poststratum is calculated and then estimate the coverage error for each block group or larger administrative unit in the country from these post-strata level estimate according to the fraction of each post-stratum it contains. However, the current article will deal with only this poststrata level estimation methodologies. A better estimator of omission rate is proposed in this paper with the form of affine combination of two model based classical DSE estimators. New estimator will be compared with existing DSE estimator and the estimator used in SRS. A potential source for bias introduced here, which is responsible for increasing the correlation bias in the working CD estimator, cannot affect our proposed estimator significantly.

2 Dual System Method

PES is done independently from census and selected households in the PES sample are checked against the census to estimate the true populatin size and omission. Let us consider a closed population

*Bayesian and Interdisciplinary Research Unit, Indian Statistical Institute, 203, B.T. Road, Kolkata-700108, India.
†E-mail: *kiranmoy07@gmail.com*

of a post-stratum U with its population size N and all individuals are homogeneous with respect to capture probabilities. Each cluster has on an average T individuals under U so that $MT = N$. Random sample of m clusters is selected for PES. From the selected clusters, individuals in 1st list(made from census) are matched one-by-one with the list made from second system - PES. The number of missed

Table 1: Distribution of sampled individuals under the population U with their observed sizes and capture probabilities in () from the Dual-System Model.

	Census Enumeration		
P-sample Population	In	out	Total
In	$x_{11}(p_{11})$	$x_{12}(p_{12})$	$x_{10}(p_{10})$
Out	$x_{21}(p_{21})$	$x_{22}(p_{22})$	$x_{20}(p_{20})$
Total	$x_{01}(p_{01})$	$x_{02}(p_{02})$	$x_{00}(p_{00})$

individuals by both systems is denoted as x_{22} and this quantity is unknown. Hence DSE estimate of N will be $\frac{M}{m}\frac{x_{10}.x_{01}}{x_{11}}$. Detailed accounts of the this coverage error estimation method along with other dual system methods are revisited in Arni et. al.(2009, [1]). Unfortunately, each of all ingredients use to calculate $\hat{N}$ is subject to error(Stark 1998, [12]). Positive dependence between two lists leads the population as under estimated(Wolter 1986, [13]). Major drawback of DSE is due to *correlation bias* which is occurred by the failure of *causal independence* and *homogeneity* in the capture probabilities within post-strata.

Suppose census actually counted C number of persons in the population then (N-C) would be the size of total omitted people. The census coverage error in terms of *net omission rate* is defined as $r = \frac{N-C}{C}$. India evaluates her census counting by estimating the omission rate corresponds to *h-th* post-stratum as, $\hat{r}_h = \frac{\hat{Y}_h}{\hat{Z}_h}$, where $\hat{Z}_h$ and $\hat{Y}_h$ are estimates of total number of enumerated and omitted persons in the *h-th* stratum respectively. Finally, India publish the omission rates only at national and zonal levels by age, sex and residence(Post Enumeration Survey 2001, published by ORGI, Govt. of India[11]). Understanding the capability of this estimate is our primary concern in this article.

3 New potential Source of Bias

At the time of PES, more efficient and well-trained enumerators(than census time) are appointed. Aim of this action is to catch more and more persons, especially to capture those people who had small chances to be captured in the census. But in some cases, this fact may raise the dependency between two systems. This causal dependence can be judge by deviation from cross-product ratio, $\theta_0 = p_{11}p_{22}/p_{12}p_{21}$ from 1. Let us define $p_{1|1}$ as the probability of the event that a person is detected at PES time when he/she was already captured by census and $p_{1|2}$ as the probability that a person is detected at PES time when he/she was not captured by census enumerator. One important part of the bias, called *correlation bias*(CB), occurs due to failure of a general independence assumption that underlies the DSEs(Griffin 2008, [6]). We decompose the bias in $\hat{r}_h$ for *h-th* post-stratum into two parts - Sampling Bias and Correlation Bias in mathematical form in main paper. Now, we can demonstrate some hypothetical situations through examples which include most of all kinds of possibilities. In human dual coverage system, $p_{1|2} < p_{1|1}$ is likely to occur i.e. census captured people are likely to be recaptured at PES time. In main paper we have taken this type of population, called *Population A*.

But for some section of the population, people are less interested to be captured second time. It may happen mostly in urban area. So, for this people $p_{1|2} > p_{1|1}$.

Table 2: Four different situations of which all have $p_{1|2} > p_{1|1}$. Exact Correlation Bias(CB) and Gross Bias of $\hat{r}$ are shown in each case. Here T= 30.

Sl. No.	$p_{1\|1}$	$p_{1\|2}$	p_{01}	CB	Bias(m=30)	Bias(m=5)
B5	0.65	0.82	0.7	0.1121	0.1134	0.1200
			0.85	0.0462	0.0466	0.0488
B6	0.65	0.90	0.7	0.1648	0.1663	0.1735
			0.85	0.0679	0.0684	0.0708
B7	0.68	0.76	0.7	0.0504	0.0515	0.0571
			0.85	0.0208	0.0211	0.0230
B8	0.70	0.90	0.7	0.1224	0.1237	0.1299
			0.85	0.0504	0.0508	0.0530

In this abstract we show only later type of population, called *Population B*. Very brief understanding from Table 2 is that for some situations estimate $\hat{r}$ is seriously affected by correlation bias as the case from B5 to B6 or B7 to B8. Hence, it is clear that conducting much more efficient enumeration at PES time(than census) may sometime distort the result under the assumption of independence in model. p_{10} increases and p_{22} tends to 0. Hence $\theta_0 \to 0$ and appropriateness of $\hat{r}$ is loosing. B-type population may be affected more by this incident. For first kind of population also, this kind problem may happen.

4 DSE-type estimators and proposed affine combination

In DSE, C and N can be estimated as $(M/m)x_{01}$ and

$$\hat{N}_\theta \;=\; (M/m)\left(\frac{x_{01}x_{10}}{x_{11}} + (\theta-1)\frac{x_{21}x_{12}}{x_{11}}\right) \tag{1}$$

for $\theta \in [0, \infty)$. Then in general $\hat{r}_\theta = \frac{\hat{N}_\theta}{\hat{C}} - 1$. The CD estimator $\hat{N}_{\theta=1} = (M/m)(x_{01}x_{10}/x_{11})$ and $\hat{r}_{(1)} = \hat{r}_{\theta=1} = x_{12}/x_{11}$. Another estimator used by Indian SRS (Raj 1977, [10]) is $\hat{N}_{\theta=0} = (M/m)x_0$, where $x_0 = x_{11} + x_{12} + x_{21}$ and then $\hat{r}_{(2)} = \hat{r}_{\theta=0} = x_{12}/x_{01}$. We consider an affine combination of the working estimator $\hat{r}_{(1)}$ and the $\hat{r}_{(2)}$ as

$$\hat{r}_u \;=\; \omega_n^*\hat{r}_{(1)} + (1-\omega_n^*)\hat{r}_{(2)} \tag{2}$$

, where $\omega_n^* \in \mathbb{R}$ and the weight ω_n^* is to be estimated. The proposed approach follows the actual underlying process and itself will decide the extent of weight should be given to the working estimator $\hat{r}_{(1)}$. The large sample approximation to the bias and variance of $\hat{r}_\theta$ are calculated and denoted by $B(\hat{r}_\theta)$ and $V(\hat{r}_\theta)$ respectively.

estimation of θ_0**.** After an extensive literature survey, a monte carlo comparison study between three chosen odds ratio estimation methods - modified mle[mmle] due to [7] and [5], median unbiased estimator[midp](Parzen et. al. 2002, [9]) and Jewell′s(1986, [8]) small sample estimator[small] has been carried out via simulation over 5000 replications. We notice that if the data (x_{11}, x_{12}, x_{21}, x_{22}) we have from those small number of subsampled clusters is almost correctly known, then $Jewell's$

small sample estimator performs significantly better than *midp* and *mmle* for both p_{01}=0.70 and 0.85. Hence we suggest small subsample size 3 is very good. Our natural aim would be to found accurate value of x_{22} for only those very small number of subsampled clusters such that atleast no person having the characteristic(neither counted in census nor included in PES list) is omitted.

Indeed, variance is not as much important as bias here. So, we shall fix an upper bound u_0 for absolute bias of $\hat{r}_u$ at our desired level and then determine simply the optimal weight by minimizing the variance $V_{\omega_n^*}(\hat{r}_u)$ with respect ω_n^* over the domain $\Omega_{u_0} = \{\omega_n^* \in \mathbb{R}^+ : |B_{\omega_n^*}(\hat{r}_u)| \leq u_0\}$. From the expressions, $B_{\omega_n^*}(\hat{r}_u)$ and $V_{\omega_n^*}(\hat{r}_u)$ are respectively estimated as $\widehat{B_{\omega_n^*}}(\hat{r}_u)$ and $\widehat{V_{\omega_n^*}}(\hat{r}_u)$ obtained by replacing maximum likelihood estimates of the parameters and $\hat{\theta}_0^s$.

Table 3: Comparison of proposed and classical dse estimators for omission rate in *Population B* based on Monte Carlo estimates of bias, variance and MSE. Here $p_{1|2} > p_{1|1}$ and m=30, T= 30, $u_0 = 0.001$.

.

	$p_{01} = 0.70$				$p_{01} = 0.85$			
	Sl. No.				Sl. No.			
	B5	B6	B7	B8	B5	B6	B7	B8
	Bias				*Bias*			
$\hat{r}_{(1)}$	0.11266	0.16606	0.05065	0.12392	0.04671	0.06833	0.02002	0.05082
$\hat{r}_{(2)}$	-0.07749	-0.04274	-0.10343	-0.04242	-0.03170	-0.01780	-0.04348	-0.01768
$\hat{r}_u$	-0.00098	-0.00089	-0.00099	-0.00087	-0.00092	-0.00071	-0.00096	-0.00071
	Variance[a]				*Variance*[a]			
$\hat{r}_{(1)}$	0.2044	0.2321	0.1662	0.1954	0.0550	0.0615	0.0456	0.0523
$\hat{r}_{(2)}$	0.0755	0.0849	0.0687	0.0854	0.0217	0.0241	0.0199	0.0241
$\hat{r}_u$	0.1078	0.0967	0.1269	0.0972	0.0325	0.0290	0.0371	0.0291
	MSE[a]				*MSE*[a]			
$\hat{r}_{(1)}$	1.4736	2.9897	0.4227	1.7310	0.2732	0.5284	0.0857	0.3106
$\hat{r}_{(2)}$	0.6760	0.2676	1.1385	0.2653	0.1222	0.0558	0.2090	0.0554
$\hat{r}_u$	0.1079	0.0968	0.1270	0.0973	0.0326	0.0290	0.0372	0.0291

[a]Numerical figures are presented in the scale of 10^{-2}.

5 Results and Conclusion

The working estimator $\hat{r}_{(1)}$ is biased and affected seriously by the dominating correlation bias factor as underlying cross-product ratio is being far from 1. By construction, estimator $\hat{r}_u$ earns a great amount of accuracy over two existing estimators $\hat{r}_{(1)}$ and $\hat{r}_{(2)}$. For *Population B*, $\hat{r}_{(2)}$ is more appropriate than $\hat{r}_{(1)}$. However, our proposed affine combination approach performs significantly better than the working estimator $\hat{r}_{(1)}$ and SRS estimator $\hat{r}_{(2)}$ in terms of accuracy and efficiency(through MSE) for both populations *Population A* and *Population B*. The analyses with Population A have been carried out in our complete study. New approach helps to get rid of from the possible problem due to unwanted hike in the dependency or correlation bias. This databased affine combination approach actually follows the true underlying process and helps to increase the efficiency by making a trade-off with accuracy level within its given bound. These results are well expected from the rigorous understanding of the nature of $\hat{r}_\theta, \theta \in [0, \infty)$ and flexible construction of $\hat{r}_u$.

References

[1] Arni S. R. Srinivasa Rao, Kiranmoy Chatterjee, B. N. Bhattacharya and Ashish Bose (2009), *Post-Enumeration Survey for the Indian Census: A Methodological Perspective*, Project Report, Indian Statistical Institute, Kolkata.

[2] Alho, J.M. and Spencer, B.D. (2005), *Statistical Demography and Forecasting*, Springer Science+Business Media, Inc.

[3] ChandraSekar, C. and Deming, W.E. (1949), *On a method of estimating birth and death rates and the extent of registration*, JASA, 44, 101-115.

[4] Freedman, D.A. and Watcher, K.A.(1994), *Heterogeneity and Census Adjustment for the Intercensal Base*, Statistical Science, 9, 476-485.

[5] Gart, J. J. and Zweifel, J. R. (1967), *On the Bias of Various Estimators of the Logit and its Variance With Application to Quantal Bioassay*, Biometrika, 54, 181-187.

[6] Griffin, R. A.(2008), *Correlation Bias Adjustment by Individual Year of Age*, DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-19, U.S. Bureau of the Census.

[7] Haldane, J. B. S. (1956), *The Estimation and Significance of the Logarithm of a Ratio Frequencies*, Annals of Human Genetics, 20, 309-311.

[8] Jewell, N. P. (1986), *On the Bias of Commonly Used Measures of Association for 2 x 2 Tables*, Biometrics, 42, 351-358.

[9] Parzen, M., Lipsitz, S., Ibrahim, J. and Klar, N.(2002), *An Estimate of the Odds Ratio That Always Exists*, Journal of Computational and Graphical Statistics, 11, 420436.

[10] Raj, D.(1977), *On Estimating the Number of Vital Events in Demographic Surveys*, JASA, 72, 377-381.

[11] *Report on Post Enumeration Survey, 2001* published by Office of Registrar General of India in June, 2006.

[12] Stark, P.B.(1998), *A Statistician's Perspective on Census Adjustment*, Presented at the Berkeley Breakfast Club, 4 December 1998.

[13] Wolter, K. M. (1986), *Some Coverage Error Models for Census Data*, JASA, 81, 338-346.