International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 22(1), 2022, pp 243-256

Some Aspects of Sampling for a Finite Population in Big Data Analysis

Md. Ayub Ali[1], Dulal Chandra Roy[2*] and Papia Sultana[2]

[1]Research Fellow, Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh

[2]Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh

[*]Correspondence should be addressed to Dulal Chandra Roy (Email: dulalroystat@yahoo.com)

[Received January 27, 2022; Accepted March 30, 2022]

Abstract

In the event of analysing big data for finite population inference, it is highly required to address issues related to selection bias (SB). In this paper, inverse sampling method was used with a view to reducing the SB related to big data sample. This method is generally known as probability proportional to size sampling (PPS) which uses the idea of auxiliary information from external sources. Results from limited simulation studies are presented. It is observed that inverse sampling method is unbiased and has better coverage rate than their alternatives (naive and calibration).

Keywords: Inverse sampling, non-probability sample, selection bias.

AMS Subject Classification: 62D05, 62R07.

1. Introduction

In the field of data science and data mining it is essential to choose appropriate probability sampling methods. Finite populating sampling is a method of drawing inference about the characteristic of a population by observing a part of the population. Probability samplings are scientific techniques in which the researcher chooses representative samples from a larger finite population based on the theory of probability. Generally, a probability sample conceives of the property that every element in the finite population has a known and nonzero probability of

being selected. Probability sampling can be employed to construct useful statistical inference of finite population characteristics. Survey sampling is an area of statistics that deals with making inference by using efficient probability sampling designs. Classical techniques in survey sampling are discussed by various authors such as Cochran (1977), Särndal et al. (1992) and Fuller (2009).

Despite the merits of probability samples, Baker et al. (2013) dispute of becoming the familiar of having non-probability samples, that may not properly represent the target population or study population. Sometimes it is pointed that some unavoidable facts such as, non-response and frame under-coverage badly hamper to take a representative non-deterministic samples in certain fields. From this point of view the prevalence of non-probability sampling techniques are more efficient compared to the other existing methods. In this circumstances, the challenges and difficulties of using non-probability samples for drawing inference for the population parameters is first introduced by Louis (2016). Elliott and Valliant (2017) used the weighting methods for reducing the selection bias (SB) in finite population inference. The process of combining information from survey data and big data using nearest neighbour imputation technique is discussed by River (2007). Bethlehem (2006) discussed sample matching methods for handling non-probability samples.

In the arena of data science and data mining it is well known that big data is one type of non- probability sample. In fact, the four significant characteristics known as Vs (Volume, Variety, Velocity and Veracity) of big data and statistical inference are also well explained in the writing of Franke et al. (2016). Now-a-days the use of big data for the purpose of predictive analysis is a very significant area of research but unfortunately analysing big data for predictive analysis in a finite population inference is not thoroughly explored in the literature. Almost all statistical fields and data science have a statistical framework to analyse big data specifically in computer science, agricultural statistics and neural networking. Tam and Kim (2018) explained the ethical challenges of big data for professional statisticians and discussed some preliminary methods of reducing for SB in big data sample. In spite of being so many challenges and difficulties of analysing big data sample, it is useful to decrease SB in the survey sampling. There are some popular methods of reducing SB related to the population parameter estimation for finite population inference.

In this paper, we have demonstrated a process using inverse sampling technique in harnessing big data for finite population inference. By negotiating the SB in the big data sample as a missing data problem, we have applied inverse sampling technique which is a special case of two-phase sampling and obtained a representative sample from the big data. The results obtained from Monte Carlo simulation studies are compared with two existing methods (e.g., naive method and calibration method) in terms of Monte Carlo bias (Bias), standard errors (SE) relative bias of the standard errors (RB SE) and coverage rate (CR).

2. Preliminaries

Let us suppose that a finite population $\{y_i : i \in U\}$, where y_i is the i^{th} observation of the study variable Y, and $U = \{1,2,3,..., N\}$ is the corresponding index set with known size N. A big data sample $\{y_i : i \in B\}$ is accessible with $B \subset U$. Specifically, $\delta_i = 1$ if $i \in B$ and $\delta_i = 0$ otherwise, and assume that y_i is observed only when $\delta_i = 1$. We are interested in estimating the population mean

$$\bar{Y}_N = N^{-1}\sum_{i=1}^{N} y_i .$$

From the big data sample B, we can estimate $\bar{Y}_N$ by $\bar{Y}_B = N_B^{-1}\sum_{i=1}^{N}\delta_i y_i$, where $\mathrm{N_B} = \sum_{i=1}^{N}\delta_i$ is the known size of B. Given $\{\delta_i : i \in U\}$, the error of $\bar{Y}_\mathrm{B}$ can be written as

$$\bar{Y}_B - \bar{Y}_N = \frac{1}{f_B} Cov(\delta, Y) \quad \text{where} \quad f_B = \frac{N_B}{N} \quad \text{and}$$

$$Cov(\delta, Y) = \frac{1}{N}\sum_{i=1}^{N}(\delta_i - \bar{\delta}_N)(y_i - \bar{Y}_N)$$

with $\bar{\delta}_N = \frac{1}{N}\sum_{i=1}^{N}\delta_i$ and thus, we have

$$E_\delta\{(\bar{Y}_B - \bar{Y}_N)^2\} = \frac{1}{f_B^2} E_\delta\{Cov(\delta, Y)^2\} \qquad (2.1)$$

where $E_\delta(.)$ denotes the expectation with respect to the random mechanism for δ_i.

If the random mechanism for δ_i is based on Bernoulli sampling, where the inclusion indicators follow Bernoulli distribution with success probability f_B independently, we can obtain

$$\begin{aligned} E_\delta\{Cov(\delta,Y)^2\} &= [E_\delta\{Cov(\delta,Y)\}]^2 + Var_\delta\{Cov(\delta,Y)\} \\ &= 0 + \frac{1}{N^2}\sum_i^N (y_i - \bar{Y}_N)^2 f_B(1-f_B) \\ &= \frac{1}{N} f_B(1-f_B)\sigma^2 \end{aligned}$$

with $\sigma^2 = N^{-1}\sum_{i=1}^N (y_i - \bar{Y}_N)^2$. Thus, under Bernoulli sampling, (2.1) reduces to

$$E_\delta\{(\bar{Y}_B - \bar{Y}_N)^2\} = \frac{1}{N_B}(1-f_B)\sigma^2$$

which is consistent with the classical theory for Bernoulli sampling with sample size $n = N_B$. For general cases, (2.1) can be expressed as

$$\begin{aligned} E_\delta\{(\bar{Y}_B - \bar{Y}_N)^2\} &= \frac{1}{f_B^2} E_\delta(Corr(\delta,Y)^2 Var(\delta)Var(Y)) \\ &= E_\delta\{Corr(\delta,Y)^2\} \times (\frac{1}{f_B} - 1) \times \sigma^2 \qquad \text{...} \qquad (2.2) \end{aligned}$$

where the second equality follows from $Var(\delta) = \frac{1}{N}\sum_{i=1}^N (\delta_i - \bar{\delta}_N)^2 = f_B(1-f_B)$

Equality (2.2) is also presented in Meng (2018). Although there are three terms in (2.2) determining the SB of $\bar{Y}_N$, the first term, $E_\delta\{Corr(\delta,Y)^2\}$, is the most critical one. Meng (2018) indicates the term as Data Defect Index (DDI), which regulates the level of departure from simple random sampling (SRS). Under equal probability sampling designs such that $E_\delta(\delta_i) = f_B$, we have $E_\delta\{Corr(\delta,Y)\} = 0$

and DDI is of order O(1/N), which implies $E_\delta\{(\bar{Y}_B - \bar{Y}_N)^2\} = \frac{1}{N}(\frac{N}{N_B} - 1)$. For other sampling designs with $E_\delta\{Corr(\delta, Y) \neq 0$, the DDI becomes significant with order O(1), which implies $E_\delta\{(\bar{Y}_B - \bar{Y}_N)^2\} = O(\frac{N}{N_B} - 1)$. Thus, a deterministic sampling design with $E_\delta\{Corr(\delta, Y) \neq 0$ marks the investigation results focus to SB.

3. Inverse Sampling Technique

We have recommended inverse sampling technique to the big data sample to correct SB. The inverse sampling can be treated as a special case of two-phase sampling (e.g., Breidt and Fuller, 1993; Rao and Sitter, 1995; Hidiroglou, 2001; Kim, et al. 2006; Stukel and Kott, 1996). The first-phase sample relates to the big data itself, which is subject to SB. The 2nd phase sample is a subsample of the first-phase sample to correct the SB of the big data sample. Hinkins et al. (1977) and Rao et al.(2003) mentioned inverse sampling technique for some usual classical sampling design, such as stratified sampling. . The application of inverse sampling to the big data subject to SB is addressed here.

In this setup, from two-phase sampling, the 1st phase sample is the big data itself and it is beyond our control. So, initially determining the SB some external sources were exploited. The next step is to find important weights for each element in the big data sample. Then select the 2nd phase sample from the big data with probability proportional to size/weights. To correct the SB using inverse sampling method, we need some external information about the target population, either from a census or from a probability sample, for some auxiliary variable x. For this, suppose that (x_i, y_i) is available in the big data sample (B) and $f(x)$ be the density for the marginal distribution of x which is taken from an external source. We have assumed x has a finite second moment. We want to estimate only $\theta = E(Y)$ using big data sample B. The first-order inclusion probability is not known for the big data sample B.

Using idea of importance sampling (Goffinet and Wallach, 1996; Henmi et al. 2007), it can be shown that

$$\hat{\theta}_{B1} = \frac{\sum_{i\in B} \frac{f(x_i)}{f(x_i|\delta_i=1)} \frac{f(y_i|x_i)}{f(y_i|x_i,\delta_i=1)} y_i}{\sum_{i\in B} \frac{f(x_i)}{f(x_i|\delta_i=1)} \frac{f(y_i|x_i)}{f(y_i|x_i,\delta_i=1)}} \tag{3.1}$$

which is asymptotically unbiased for $\theta = E(Y)$ by considering that $f(\delta_i = 1|\, x_i) > 0$ for $i \in U$ almost surely. If the sampling mechanism for big data sample B is ignored after controlling on x, i.e. $P(\delta_i = 1|\, x_i, y_i) = P(\delta_i = 1|\, x_i)$, then (3.1) reduces to

$$\hat{\theta}_{B1} = \frac{\sum_{i\in B} \frac{f(x_i)}{f(x_i|\delta_i=1)} y_i}{\sum_{i\in B} \frac{f(x_i)}{f(x_i|\delta_i=1)}} = \sum_{i\in B} w_{i1} y_i \tag{3.2}$$

Here weights w_{i1} can be called as an importance weight, following the idea of importance sampling. If x_i is the vector of stratum indicator variables, then $\frac{f(x_i)}{f(x_i|\delta_i=1)}$ equals to $\frac{(N_h/N)}{(n_h/n)}$ for i in stratum h, and this leads to unbiased estimation under stratified sampling.

If only $\bar{X}_N = \frac{\sum_{i=1}^{N} x_i}{N}$ is available, we can approximate $f(x)$ by $f_0(x)$, which minimizes the Kullback-Leibler distance

$$\min_{f_0 \in P_0} \int f_0(x) \ln\{\frac{f_0(x)}{f(x|\,\delta=1)}\} dx \tag{3.3}$$

where $P_0 = \{f(x); \int x f(x) dx = \bar{X}_N\}$.

The solution to (3.3) is

$$f_0(x) = f(x|\,\delta=1) \frac{\exp(x^T\lambda)}{E\{\exp(X^T\lambda|\,\delta=1)\}} \tag{3.4}$$

where λ satisfies $\int x f_0(x) dx = \bar{X}_N$, and D^T is the transpose of D. Therefore, the selection probability for the second-phase selection is proportional to $\exp(x^T\lambda)$, which is very close to the exponential tilting calibration discussed in Kim (2010). By applying (3.4), the weighted estimator in (3.2) became

$$\hat{\theta}_{B1} = \frac{\sum_{i\in B} \exp(x_i^T \hat{\lambda}) y_i}{\sum_{i\in B} \exp(x_i^T \hat{\lambda})} \tag{3.5}$$

where $\hat{\lambda}$ satisfies $$\frac{\sum_{i\in B} \exp(x_i^T \hat{\lambda}) x_i}{\sum_{i\in B} \exp(x_i^T \hat{\lambda})} = \bar{X}_N \tag{3.6}$$

Here, equation (3.6) is known as calibration equation (Wu and Sitter, 2001). We may ignore the sampling variability in estimating the parameter λ contrasting the usual calibration estimation since N_B is large. It will be challenging to solve calibration equation (3.6) when the sample size of B is large. Hence, one-step approximation (Kim, 2010) can be used.

Based on (3.5), we can demonstrate how the 2nd phase sample (B_2) of size n can be taken from the big data sample B such that $\hat{\theta}_{B2} = \frac{\sum_{i=1}^{B2} y_i}{n}$ is approximately design unbiased for $\hat{\theta}_{B1}$ in (3.2). Therefore, the main idea is to take the conditional first-order inclusion probability $\pi_{i2|1} = P(i \in B_2 \mid i \in B)$ such that

$$\pi_{i2|1} = n w_{i1},\ i \in B \tag{3.7}$$

where w_{i1} is the importance weight in (3.2). To assure

$$\pi_{i2|1} \in (0,1], i \in B \tag{3.8}$$

we should select $n \le \frac{1}{\max_{i\in B}\{w_{i1}\}}$. Once $\pi_{i2|1} : i \in B$ satisfying (3.7) and (3.8) are found, we can apply any unequal probability sampling methods to get a 2nd phase sample; see Tille (2006) for details on algorithms for unequal probability sampling designs.

After selecting 2nd phase sample B_2, we use the sample mean of y_i in B_2 to estimate θ and the variance estimator of $\hat{\theta}_{B2}$ can be found as

$$Var(\hat{\theta}_{B2}) = Var(\hat{\theta}_{B1}) + Var(\hat{\theta}_{B2} - \hat{\theta}_{B1})$$

where the first term is of order $O(\frac{1}{N_B})$, and the second term is of order $O(\frac{1}{n})$. If $\frac{n}{N_B} = o(1)$, the first term can be safely ignored, and we only need to estimate the second term. Since we can express

$$\hat{\theta}_{B2} = \sum_{i \in B_2} \frac{1}{\pi_{i2|1}} (w_{i1} y_i),$$

we can apply the standard variance estimation formula for the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) by treating the big data as the finite population.

That is, we can use

$$\hat{V} = \sum_{i \in B_2} \sum_{j \in B_2} (\frac{\pi_{ij2|1} - \pi_{i2|1}\pi_{j2|1}}{\pi_{ij2|1}})(\frac{w_{i1} y_i}{\pi_{i2|1}})(\frac{w_{j1} y_j}{\pi_{j2|1}})$$

which is a variance estimator for $\hat{\theta}_{B2}$, where $\pi_{ij2|1}$ is the joint inclusion probability for the 2nd phase sample.

4. Simulation Study

For the simulation study, we use inverse sampling method under a simple setup. Our finite population is generated by using the following model:

$$y_i = 4 + 2x_i + e_i, \qquad i = 1,2,..., N,$$

where $x_i \sim \text{Exp}(2), e_i \sim N(0, x_i^2), N = 1{,}000{,}000, N(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2, and $Exp(\lambda)$ is an exponential distribution with mean λ. The inclusion indication of the big data sample is generated by

$\delta_i \sim Ber(p_i)$ independently for $i = 1,2,..., N$,where $\mathrm{p_i} = \sqrt{2\pi} \times \phi(\mathrm{x_i} - 2)$, and a standard normal function ϕ which creates probabilities proportional to their sizes.

The selection probability for the second-phase selection is proportional to $\exp(x^T \lambda)$, which is very close to the exponential tilting calibration discussed in Kim (2010). Thus, the weighted estimator became $\hat{\theta}_{B1} = \frac{\sum_{i \in B} \exp(x_i^T \hat{\lambda}) y_i}{\sum_{i \in B} \exp(x_i^T \hat{\lambda})}$ where $\hat{\lambda}$ satisfies $\frac{\sum_{i \in B} \exp(x_i^T \hat{\lambda}) x_i}{\sum_{i \in B} \exp(x_i^T \hat{\lambda})} = \bar{X}_N$

In addition, we assume that the population mean $\bar{X}_N$ is known. Considering two cases of ϕ function, we are interested in making inference for the population mean $\bar{Y}_N$ and a proportion, where $\mathrm{I(x < 1)} = 1$ if $x < 1$ for a given number a=1, and 0 otherwise.

We compute the following three estimators with $n = 600$ and $n = 1200$, respectively, and recall that n is the sample size for the 2nd phase sample.

I. **Naive estimator:** For getting a sample of size n we use simple random sampling without replacement (SRSWOR) from the big data sample B.

II. **Calibration estimator:** After taking the sample obtained by the naive method, we apply exponential tilting method mentioned above and obtain a calibration estimator in terms of $\bar{X}_N$ information.

III. **Inverse sampling estimator:** Considering importance weights using (3.5) satisfying the calibration condition (3.6), and then a sample of size n is taken by probability proportional to size sampling (PPS).

Here we compute 500 Monte Carlo (MC) simulations for estimating population mean and proportion by using naive, calibration and inverse sampling method. We have compared these methods in terms of the bias and standard errors (SE), the relative bias (RB) and coverage rate of a 95% confidence interval (CI) using the Wald-type method.

First, we have presented the population data (N=10,00,000) in the following graph for the visualization.

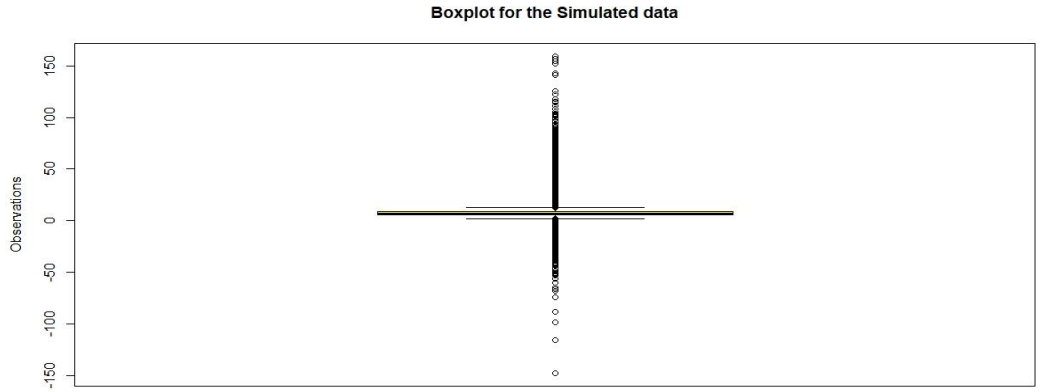

Figure 4.1: Data visualization using box plot

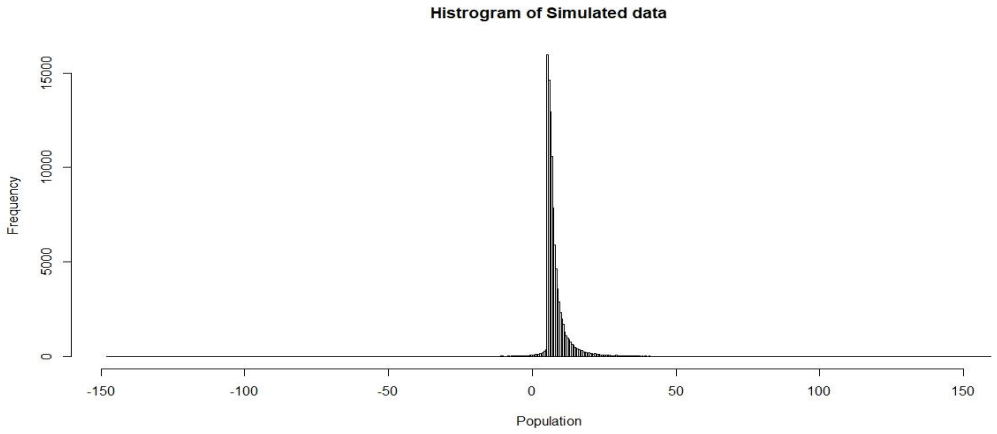

Figure 4.2: Data visualization using histogram

The data visualization given in the above diagram 4.1 (box plot) and 4.2 (histogram) shows that there is a heterogeneity in the simulated dataset. So from this heterogeneity characteristic we can easily use probability proportional to size sampling (PPS) and estimate the population parameters, mean and proportion. This will help us to draw inference for the finite population. The table below shows the Monte Carlo simulation results:

Table 4.1: Bias, SE, RB.SE and CR for the different estimators

Par.	Method	ϕ	Bias	SE **n=600**	RB.SE	CR	Bias	SE **n=1200**	RB.SE	CR
$\bar{Y}_N$	Naiv.	-0.3	-0.251	0.125	0.026	0.440	-0.302	0.095	-0.019	0.323
	Cal.		0.015	0.070	-0.014	0.941	0.011	0.054	-0.018	0.932
	Inv.		0.000	0.140	0.015	0.950	0.000	0.110	0.000	0.960
	Naiv.	-0.6	-0.491	0.112	0.022	0.201	-0.495	0.083	0.000	0.000
	Cal.		0.015	0.080	-0.051	0.943	0.014	0.057	-0.029	0.934
	Inv.		0.000	0.150	0.000	0.950	0.000	0.120	0.001	0.950
P_N	Naiv.	-0.3	-0.022	0.022	-0.027	0.783	0.018	0.015	-0.009	0.693
	Cal.		0.012	0.022	0.028	0.932	0.013	0.013	0.017	0.891
	Inv.		0.000	0.022	0.022	0.931	0.000	0.013	0.022	0.920
	Naiv.	-0.6	-0.043	0.023	0.000	0.412	0.049	0.018	0.016	0.181
	Cal.		-0.012	0.023	0.000	0.931	-0.011	0.015	0.028	0.919
	Inv.		-0.013	0.022	0.000	0.950	-0.001	0.013	0.021	0.960

Note that Naiv., Cal. and Inv. are the short forms of naive, calibration and inverse sampling estimators respectively

To get a clearer idea, we have prepared some graphs. In all the graphs, the estimators (naive, calibration and inverse) are compared on the basis of Bias, SE, RB.SE and CR obtained from the simulation studies for different sample sizes n and for the different values of ϕ.

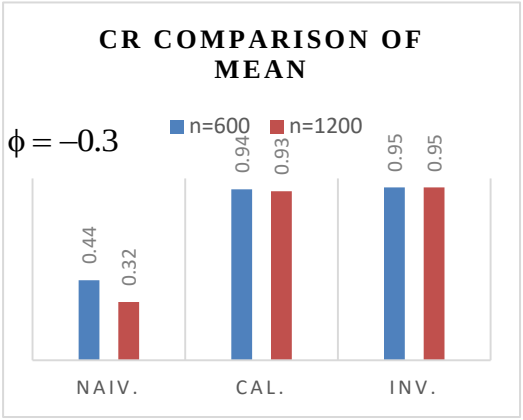

Figure: 4.3

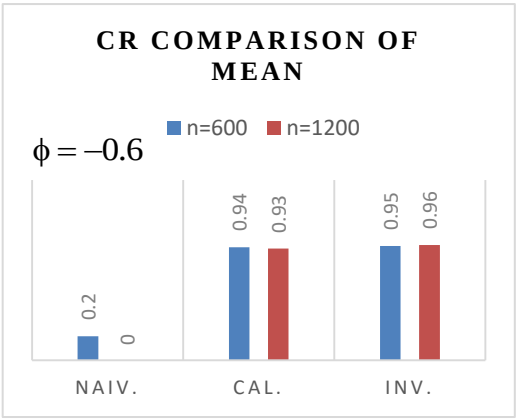

Figure: 4.4

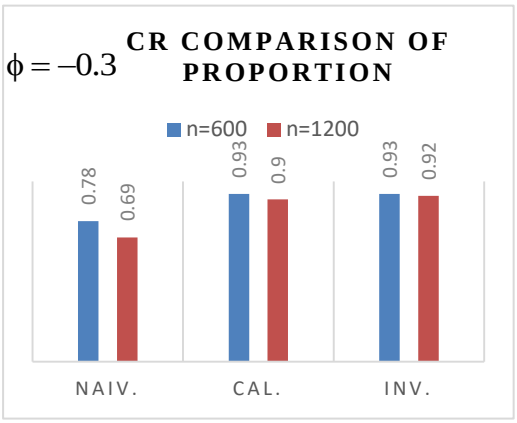

Figure: 4.5

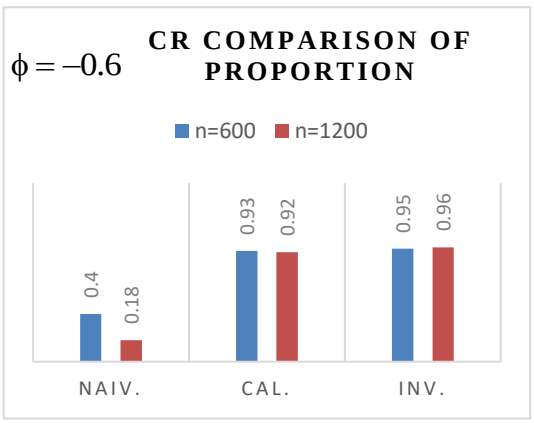

Figure: 4.6

From table 4.1 and the different graphs produced above, we find that the performance of naive estimator is very poor. Precisely, increasing size of sample creates low coverage rate for the population parameter. Although for estimating the population mean $\bar{Y}_N$ (linear function of $\bar{X}_N$), the calibration estimator works better compared to naive method but inverse sampling estimator is unbiased and has a better coverage rate (around 95%). Similarly, for population proportion P_N (not a linear function of $\bar{X}_N$), the bias and the standard errors of the calibration estimator and the inverse sampling estimator are nearly the same, but the inverse sampling estimator has a better coverage rate compared to other two methods (naive and calibration method). Finally, we can conclude that the inverse sampling method provides better results compared to naive and calibration methods.

Acknowledgement

The authors acknowledge the two anonymous reviewers and the editor for insightful comments that improved the presentation and clarifications of our manuscript. The authors are also thankful to National Science and Technology (NST), Bangladesh, for the fellowship to complete this work.

References

[1] Cochran, W. G. (1977). Sampling Techniques, 3rd edn, John Wiley & Sons, New York.

[2] Sarndal, C. E., Cassel, C. M. and Wretman, J. H. (1992). Model Assisted Survey Sampling, Springer-Verlag, New York.

[3] Fuller, W. A. (2009). Sampling Statistics, John Wiley & Sons, Hoboken.

[4] Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J. and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling, J. Surv. Stat. Methodol., 1, 90–143.

[5] Keiding, N. and Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys (with discussions), J. Roy. Statist. Soc. Ser. A, 179, 1–28.

[6] Elliott, M. and Valliant, R. (2017). Inference for non-probability samples, Stat. Sci., 32, 249–264.

[7] Rivers, D. (2007). Sampling for web surveys, Proceedings of the Survey Research Methods Section, American Statistical Association.

[8] Bethlehem, J. G., Fannie Cobben, and Barry Schouten (2006). Nonresponse in household surveys.

[9] Franke, B., Plante, J.-F., Roscher, R., Lee, E.-S. A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M. M., Grosse, R., Hendricks, D. and Reid, N. (2016). Statistical inference, learning and models in big data, Int. Stat. Rev., 84, 371–389.

[10] Tam, S.-M. and Kim, J. K. (2018). Big data, selection bias and ethics – an official statisticians's perspective, Stat. J. IAOS. Accepted for publication.

[11] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, J. Amer. Statist. Assoc., 47(260): 663–685.

[12] Goffinet, B. and Wallach, D. (1996). Optimized importance sampling quantile estimation, Biometrika , 83, 791–800.

[13] Henmi, M., Yoshida, R. and Eguchi, S. (2007). Importance sampling via the estimated sampler, Biometrika, 94, 985–991.

[14] Kim, J. K. (2010). Calibration estimation using exponential tilting in sample surveys, Surv. Methodol., 36, 145–155.

[15] Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data, J. Amer. Statist. Assoc., 96, 185–193.

[16] Tillé, Y. (2006). Sampling Algorithms, Springer-Verlag, New York.

[17] Tillé, Y. (2016). Unequal probability inverse sampling, Surv. Methodol., 42, 283–295.

[18] Breidt, F. J., and W. A. Fuller (1993). Regression weighting for multipurpose samplings. Sankhya: Ser. B 55: 297-309.

[19] Rao, Jon N. K, and R. R. Sitter (1995). Variance estimation under two-phase sampling with application to imputation for missing data. Biometrika 82.2: 453-460.

[20] Hidiroglou, M. A. (2001). Double sampling. Survey methodology 27.2: 143-154.

[21] Robotham, Hugo, Zaida I. Young, and Juan C. Saavedra-Nievas (2008). Jackknife method for estimating the variance of the age composition using two-phase sampling with an application to commercial catches of swordfish (Xiphias gladius). Fisheries research 93.1-2: 135-139.

[22] Apps, J. A., E. L. Madsen, and R. L. Hinkins (1977). The kinetics of quartz dissolution and precipitation. Geothertnal and Geosciences Progratn: 12.

[23] Subba Rao, N. (2003). Groundwater quality: focus on fluoride concentration in rural parts of Guntur district, Andhra Pradesh, India. Hydrological Sciences Journal 48.5: 835-847.