Mathematics and Statistics 7(4): 91-97, 2019 http://www.hrpub.org
DOI: 10.13189/ms.2019.070401

Assessing the Effect of Complex Survey Design in the Analysis of Child Labour Prevalence Rate in Ghana

Lucy Twumwaah Afriyie[1,*], Bashiru I. I. Saeed[2], Abukari Alhassan[1]

[1]Department of Statistics, University for Development Studies, Ghana
[2]Department of Mathematics and Statistics, Kumasi Technical University, Ghana

Received May 7, 2019; Revised August 22, 2019; Accepted August 29, 2019

Abstract Statistical surveys are conducted to estimate population parameters where there are reasons restricting the use of the total population. In practice, there are two different survey strategies (i.e. simple and complex survey designs) to be implemented and the choice of a strategy depends on several factors including the characteristics of the population, the nature of the research questions, etc. However, when the complex survey design is used, standard statistical methods that do not take into account the complex nature of the survey design may lead to inaccurate estimates. In Ghana, living standard surveys are conducted using complex survey design involving stratifications, clustering and estimation of survey weights. In this study, bootstrap resampling methods are used to explore the effect of complex survey design in the analysis of child labour prevalence rate. The relative efficiency of the complex survey design approach was determined by using design effect (*deff*). Data from the Ghana Living Standard Survey Round 6 (GLSS 6) conducted by the Ghana Statistical Service in 2012 was used for the analysis and the target population was children aged 5–17 years. The results from the simulation study shows that relative efficient estimates are obtained when the complex survey design characteristics are considered in the analysis. Thus, ignoring the characteristics of complex survey design could lead to unrealistic estimates.

Keywords Complex Survey Design, Design Effect, Relative Efficiency, Standard Error, Survey

1. Introduction

In many situations, analysis on elements in a target population may not be feasible because of different reasons including cost, time, difficulty in reaching all elements etc. [1]. In such a situation dealing with samples from the population becomes the ideal alternative. The selection of the samples from the target population for the survey can be done using different sampling techniques [1]. In most cases simple random sampling approach is typically used when the population under consideration is homogeneous. However, due to the complex nature of some population, many studies result to multi-stage sampling which involves the use of more than one approach in selecting the sample. That is, the population is divided into groups and sub-groups and then select elements at each stage. This approach is usually referred to as complex survey design (CSD).

In practice, when complex survey design approach is used, the estimates and their associated variances becomes different from what would have been obtained if simple random sampling approach is used [2, 3]. This is dues to the fact that the computation of the estimates in CSD takes into consideration the stratification of the population into sub-groups and some weights which is calculated to expand the results to reflect the population. The uses of such weights allow to compute an estimate which is unbiased of the population parameter [4]. According to Heeringa, West [5], the standard practice in the design-based analysis of complex sample survey data requires the analysts to identify variables containing final survey weights to cater for unequal probabilities of selection, nonresponse and also post-stratification adjustments. However, many studies use the CSD approach but ignore the use of the weight during the computation of the estimates. A review by the American Public Health Association in 2012 showed that only 60% of published articles accounted for the survey weights when computing estimates [2]. The occurrence and precision of reporting differ among publication in the various journals.

In Ghana, living standard surveys are conducted by the Ghana Statistical Services (GSS) at certain time periods to support decision making on living conditions of the populace. The living standard surveys are conducted using CSD due to variability in the different administrative regions in Ghana. Although the estimates reported by the GSS takes into consideration the complex nature of the

survey, the issue under consideration is, which of the two approaches lead to less sample-to-sample variability. Such information is important to avoid misleading inference in the parameter estimates and its uncertainty [6]. This study addresses this knowledge gap by conducting simulation study to determine which approach is more suitable in practice. That is, the study seeks to conduct a simulation study using bootstrapping methods to explore the effect of complex survey design in the analysis of child labour prevalence rate. Through the simulation study the clear difference between the variability in each approach can be determined and the relative efficiency between the two approaches can be calculated.

2. Materials and Methods

2.1. Data Description

In this study, data from the Ghana Living Standard Survey Round 6 (GLSS 6) conducted by the Ghana Statistical Service (GSS) in 2012 was used for the analysis and the target population was children aged 5–17 years [7]. A two-stage sampling design was adopted during data collection. A total of 18,000 households were selected nationwide, out of which 16,772 were finally interviewed, yielding a response rate of 93.2%. Within the 16772 households, information was independently collected for 24,116 children aged 5–17 years (sampling unit). Each sampling unit in the target population had a known, non-zero probability of being included in the sample. The 2010 Population and Housing Census (PHC) conducted by the Ghana Statistical Service were used as the sampling frame for the GLSS6. Details of the survey can be obtained from the GLSS6 report [7].

The main variables analyzed in this study were child labour (child labour/ not child labour), child's age (5–9/ 10–14/ 15 –17), child's gender (male/female), child's current grade (No basic education/ primary/ junior high school, senior high school/ tertiary), presence of the father (Yes/ No) and mother in the household (Yes/ No), relationship to the household head (son/daughter/other relative), educational grade completed by the father and mother (formal education/no formal education) and locality of residence (urban/rural).

2.2. Bootstrapping Methods

Sample from the population can provide all estimates of interest and their simulation distribution gives detailed insight in their performance. When the sampling distribution of a statistic and its characteristics are unknown, they need to be estimated or approximated from data, which can be done using bootstrap approaches [8, 9]. Bootstrapping is a technique used to obtain a description of the sampling properties of empirical estimators using the sample data themselves, rather than broad theoretical results [10-12]. The applications of bootstrapping in statistical analysis have been popular because of its ability to handle in sample surveys and also easily deal with situation that cannot be compared theoretically by analysts [13, 14]. The bootstrapping approach is widely used for statistical inference [9, 14, 15] and to measure the uncertainty associated with an estimator. The concept of bootstrapping involves repeatedly drawing independent samples from a single population with-replacement to form a distribution rather than a single sample [11, 16]. Bootstrap population would allow sampling according to the sampling plan of interest, be close to the respective national population regarding distributions of interesting variables.

Let $S = \{y_1, y_2, \ldots, y_n\}$ represent an independent sample of size n drawn from a population P and θ be a statistic of interest (*proportion of child labour in our case*) that describes P. The value of θ can be estimated by calculating $\hat{\theta}$ from the sample S [17]. The sampling distribution of $\hat{\theta}$ can be estimated through bootstrapping technique. The general routing of the bootstrapping technique is outline as follows:

1) Generate a simple random sample of size n from S with-replacement to obtain a bootstrap sample S_1^* such that each element has $\frac{1}{n}$ probability of being selected.
2) Calculate the bootstrap version of the statistic of interest, $\hat{\theta}_1^*$
3) Repeat the process in step 1 and 2 B number of times to obtain bootstrap estimates $\hat{\theta}_b^* \{b = 1, \ldots, B\}$ of the parameter of interest θ.

The estimate of the parameter and the associated standard error can be obtained as:

$$\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^* \quad (1)$$

$$SE(\hat{\theta}^*) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}^*)^2} \quad (2)$$

For a complex survey design, the stratified bootstrapping approach which takes into consideration the weights calculated based on the stratification of the sample unit within the population is used [18]. The stratified bootstrapping routing is outline as follows:

1) Generate a simple random sample of size n_h from S_h with-replacement to obtain a bootstrap sample

$S_{1h}^{\#}$ in stratum $h\ \{h=1,\ldots,H\}$, independently for each stratum.

2) Calculate the bootstrap version of the statistic of interest $\hat{\theta}_{1h}^{\#}$ in stratum h, and $\hat{\theta}_{1}^{\#}=\sum_{h=1}^{H} w_h \hat{\theta}_{1h}^{\#}$ where w_h represents the calculated stratum weight.

3) Independently repeat the process in step 1 and 2 B number of times to obtain bootstrap estimates $\hat{\theta}_{b}^{\#}\{\mathrm{b}=1,\ldots,\mathrm{B}\}$ of the parameter of interest θ.

The estimate of the parameter and the associated standard error can be obtained as:

$$\hat{\theta}^{\#}=\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{b}^{\#} \tag{3}$$

$$SE(\hat{\theta}^{\#})=\sqrt{\frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_{b}^{\#}-\hat{\theta}^{\#})^{2}} \tag{4}$$

The comparison of the two approaches of estimating the population parameter is achieved by relative efficiency measure called design effect (*deff*). The *deff* is the ratio of the variances for the two approaches:

$$deff=\frac{\sum_{b=1}^{B}(\hat{\theta}_{b}^{\#}-\hat{\theta}^{\#})^{2}}{\sum_{b=1}^{B}(\hat{\theta}_{b}^{*}-\hat{\theta}^{*})^{2}} \tag{5}$$

$Deff<1$ means the bootstrapping estimates under multi-stage sampling is relatively efficient than the estimate under the simple random sampling, $Deff>1$ gives opposite meaning whereas $Deff=1$ mean the two approaches are equally efficient.

3. Results

To investigate the effect of complex survey design in the analysis of child labour prevalence rate a Monte Carlo simulation experiment was performed. In the simulation experiment, 5,000 bootstrap replications were generated based on the two approaches described in the previous section. In the first approach 5,000 bootstrap samples were generated using simple random sampling with-replacement (referred to as unweighted) whilst in the other case we consider the complex survey design by including the stratification and the survey weights in selecting the bootstrap samples (referred to as weighted). Large bootstrap samples were used because it improves estimation performance [19]. Generally, it has been observed in the literature that, the value of the design effect differ across variables measured in the sample surveys, and estimates of design effect are typically provided for the major variables of interest but in this paper, variables related to the child labour were also considered.

The results from both experiments are presented in Table 1. In both cases, the prevalence rate of child labour and the associated standard error are presented together with the other variables considered. In general, the results show that there is no much difference in the estimated proportion of child labour by the two approaches. The variations in the percentage standard error for both approaches also confirm that there is no much difference between the two approaches. However, the little differences that exist in the two approaches favours the weighted approach in most cases. That is, the calculated *deffs* is lower than 1 in most cases suggesting that the weighted approach is relative efficient in most cases compared to the unweighted approach. For variables child's age and gender the *deffs* suggest that the unweighted approach is relative efficient than the weighted approach whereas the weighted approach remains relative efficient for all the remaining variables. This result is in agreement with the existing literature [4, 20] that concluded that variance estimation that takes into account the complex nature of the sampling plan leads to less bias estimates.

On the other hand, the general results contradict the findings by Salganik [6]. The study by Salganik [6] found that the estimates from CSD are less precise compared to simple random sample but not always. The differences in the result may be influence by the differences in sample sizes. In Salganik [6] study, the sample size used range between 69 and 374 whilst the current study used 24,116, which is sufficiently large to change the findings. The authors further recommended the use of sample size twice as large as would be needed under unweighted sampling scheme to achieve a better result.

The densities of the estimated prevalence rate from the two approaches were also compared to determine the closeness of the center of the two distributions. Figure 1 describes the densities of the distributions of the bootstrap sample for all the variables considered. Since almost all the variables are binary, the density plots were displayed for only one outcome. The results described by the figure suggest that although the estimated proportions and their associated standard error do not vary much, the center of the two distributions in most cases are distinct. For variables child labour and urban residence, the density of weighted sample does not significantly contain the center of the unweighted density and vice versa. However, for the remaining variables the two densities contain the center of each one. For variables age (12-17) and mother lived in household the two densities nearly coincided with each other.

Table 1. Simulation result on percentage distribution of estimate, standard error and *deff*

Variable	Unweighted		Weighted		*Deff*
	Proportion	Std. Error (%)	Proportion	Std. Error (%)	
Child labour					
Yes	24.28	0.28	25.35	0.27	0.97
No	75.72	0.28	74.65	0.27	0.97
Child's age					
5-9	41.19	0.32	41.45	0.31	1.00
10-14	39.03	0.31	38.77	0.31	1.05
15-17	19.78	0.26	19.78	0.26	1.01
Child's gender					
Male	51.04	0.32	50.82	0.33	1.01
Female	48.96	0.32	49.18	0.33	1.01
Child ever attend school					
Yes	91.98	0.17	92.11	0.17	0.91
No	8.02	0.17	7.89	0.17	0.91
Locality of residence					
Urban	35.29	0.31	36.03	0.28	0.83
Rural	64.71	0.31	63.97	0.28	0.83
Relationship to household head					
Son/daughter	78.34	0.27	78.25	0.26	0.96
Other relative	21.66	0.27	21.75	0.26	0.96
Father lived in household					
Yes	65.20	0.31	65.20	0.31	0.97
No	34.80	0.31	34.80	0.31	0.97
Mother lived in household					
Yes	79.11	0.26	79.06	0.26	0.96
No	20.89	0.26	20.94	0.26	0.96
Mother's education					
Formal education	8.45	0.18	8.46	0.18	0.94
No formal education	91.55	0.18	91.54	0.18	0.94
Father's education					
Formal education	18.68	0.25	18.59	0.24	0.93
No formal education	81.32	0.25	81.41	0.24	0.93

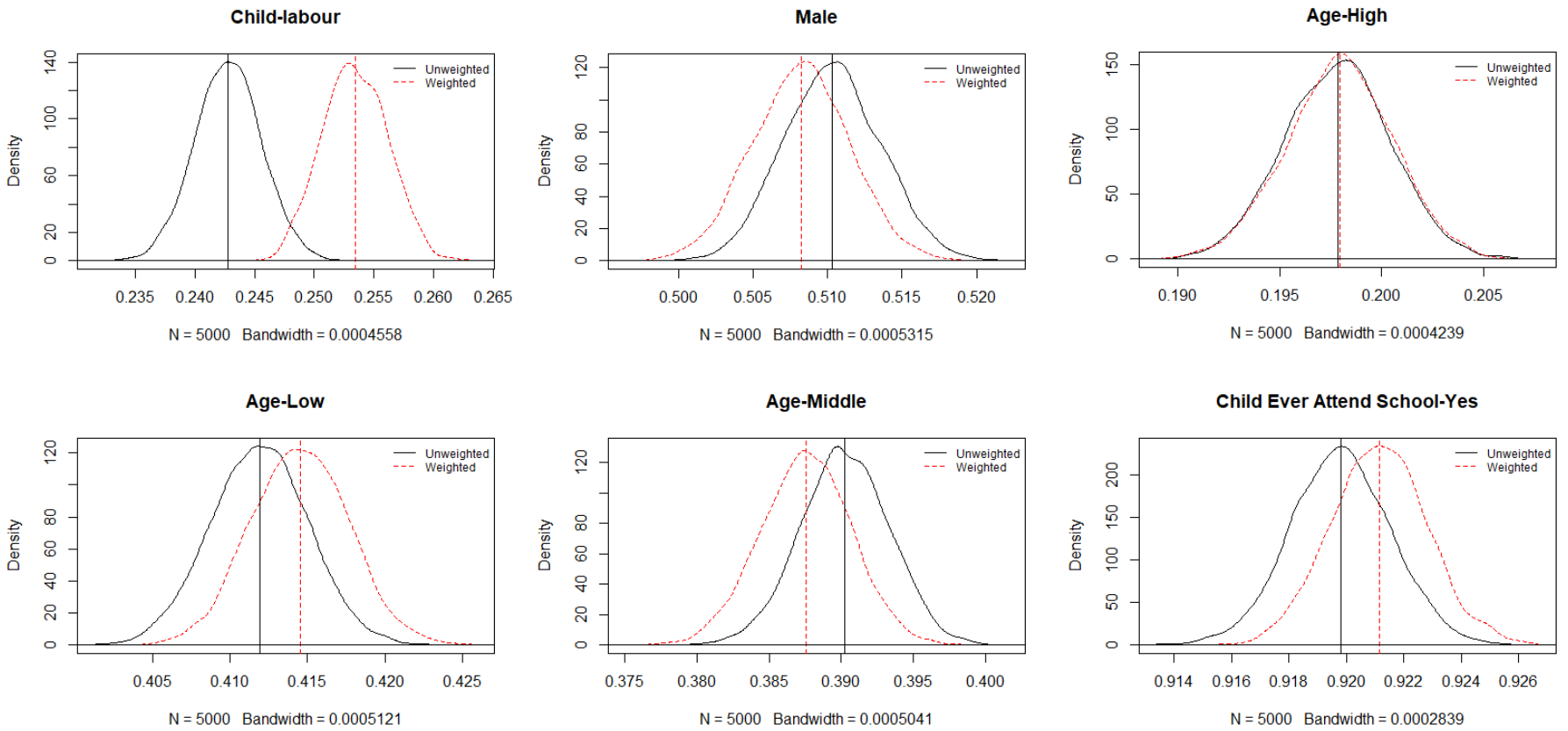

Figure 1. Density plot of the 5,000 unweighted (solid line) and weighted (broken lines) bootstrap samples generated from the GLSS6 data.

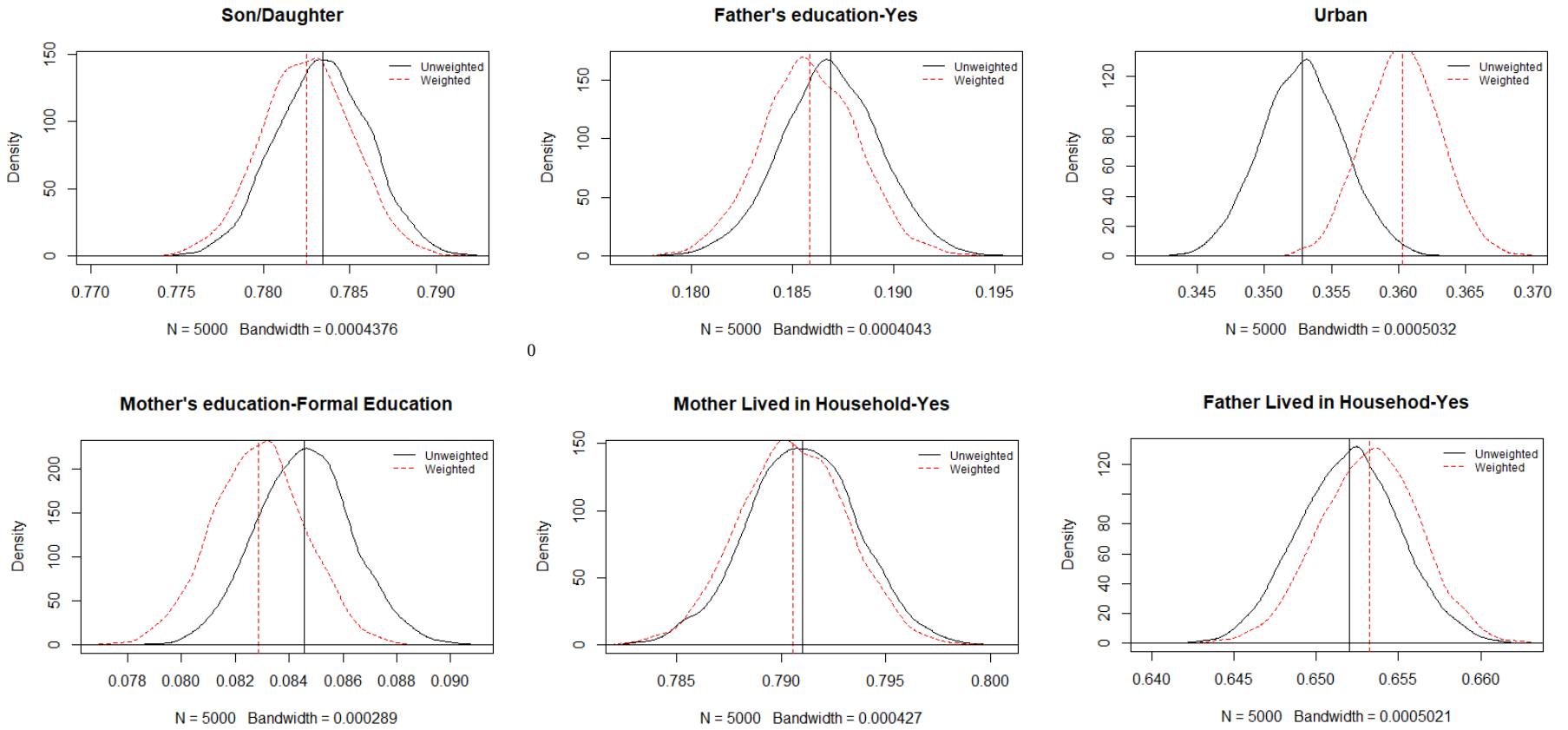

Figure 1 (Continued). Density plot of the 5,000 unweighted (solid line) and weighted (broken lines) bootstrap samples generated from the GLSS6 data.

4. Conclusions

Household surveys are commonly designed to produce estimates of population totals, means, or proportion. Examples of totals might be total population of the number of children aged 5-17 years involved in child labour. In Ghana, results from the Ghana living standard surveys Round 6 are used to support decision making processes regarding the situation under consideration. In most cases, CSD is applied rather than simple random sampling. However, when the complex data sources are not accounted for and variances are not estimated appropriately, it may leads to an increment in type I errors [3]. This has substantial implications for the way in which the survey data may be acceptable to the wider community and used in policy development [3]. Though this is the case, many studies in the literature fail to account for such variation due to the complexity of the survey [2].

In this study the effect of CSD in the analysis of child labour data were explored using bootstrapping resampling methods. In the simulation study 5,000 bootstrap samples were obtained from the data and the estimated proportion of child labour and the associated standard error were obtained. In the study two different bootstrap strategies: unweighted and weighted were considered. From the simulation study it was found that in most cases, the estimation is relative efficient when the complex data sources are taken into consideration. Although, there are differences in the results from the two approaches, the design effect suggests small difference. This result may be influenced by the fact that the sample size of the survey data considered was higher (over 24,000). Due to high sample size the variation among the two methods were not substantial.

Acknowledgements

I am most grateful to Prof. N.N.N Nsowah-Nuamah, Dr. Eric N. Aidoo (Department of Mathematics, KNUST, Ghana), and Dr. Smart Sarpong (KsTU, Ghana) for their useful comments and suggestions. I am also thankful to the management of the Ghana Statistical Service.

REFERENCES

[1] Cochran, W.G., Sampling techniques. John Wiley and Sons, New York. 1977.

[2] Bethany, A.B., et al., Use of design effects and sample weights in complex health survey data: A review of published articles using data from 3 commonly used adolescent health surveys. American Journal of Public Health, 2012. **102**(7): p. 1399-1405.

[3] Burden, A., et al., The impact of complex survey design on prevalence estimates of intakes of food groups in the Australian national children's nutrition and physical activity survey, centre for statistical and survey methodology, university of Wollongong, Working Paper 24-10, 21p. http://ro.uow.edu.au/cssmwp/74, 2010.

[4] West, B.T. and S.E. McCabe, Incorporating complex sample design effects when only final survey weights are available. Stata Journal, 2012. **12**(4): p. 718-725.

[5] Heeringa, S.G., B.T. West, and P.A. Berglund, Applied survey data analysis. Chapman and Hall, UK. 2010.

[6] Salganik, M.J., Variance estimation, design effects, and sample size calculations for respondent-driven sampling. Journal of Urban Health: Bulletin of the New York Academy of Medicine, 2006. **83**(7): p. i98-i111.

[7] Ghana Statistical Service, Ghana living standard survey (Round 6). Accra. Ghana Statistical Service. 2012.

[8] Yaquia, Z. and J. Kolassa, Assessing and comparing the accuracy of various bootstrap methods. Communications in Statistics - Simulation and Computation, 2017.

[9] Honoré, B.E. and L. Hu, Easy bootstrap-like estimation of asymptotic variances. Economics Letters, 2018. **171**: p. 46-50.

[10] Davison, A.C. and D.V. Hinkley, Bootstrap methods and their application. Cambridge University Press, UK. 1997.

[11] Efron, B. and R.J. Tibshirani, An introduction to the bootstrap. Chapman and Hall, UK. 1993.

[12] Efron, B., Nonparametric standard errors and confidence intervals (with discussion). Canadian Journal of Statistics, 1981. **9**: p. 139-172.

[13] Munnich, R. and J. Schurle, On the simulation of complex universe in the case of applying the German microcensus. The DACSEIS Research Paper Series 5, Tubingen, 2003.

[14] Haziza, D. and C. Leger, A survey of bootstrap methods in finite population sampling. Statistics Survey, 2016. **10**: p. 1-52.

[15] Wang, J., et al., Assessing local determinants of neural tube defects in the heshuns region, Shanxi Province, China. BMC Public Health, 2010. **10**(52): p. 1-11.

[16] Adhya, S., Bootstrap variance estimation for semiparametric finite population distribution function estimator. Calcutta Statistical Association Bulletin, 2018. **70**: p. 17-32.

[17] Carsey, T.M. and J.J. Harden, Monte carlo simulation and resampling methods for social science. Sage Publications Ltd, UK. 2014.

[18] Rao, J.N.K., C.F.J. Wu, and K. Yue, Some recent work on resampling methods for complex surveys. Survey Methodology, 1992. **18**: p. 209-217.

[19] Banjanovic, E.S. and J.W. Osborne, Confidence interval for effect sizes: Applying bootstrap resampling. Practical Assessment Research & Evaluation, 2016. **21**(5): p. 1-20.

[20] Wagner, H. and D. Eckmair, Simulation studies for complex sampling designs. Austrian Journal of Statistics, 2006. **35**(4): p. 419-435.