academicJournals

Vol. 6(4), pp. 58-63, April 2013
DOI 10.5897/AJMCSR 11.129
ISSN 2006-9731 ©2013 Academic Journals
http://www.academicjournals.org/AJMCSR

African Journal of Mathematics and Computer Science Research

Full Length Research Paper

Specification parameters for linear estimators in probability proportional to size sampling scheme

Ikughur, Atsua Jonathan[1]* and Amahia, Godwin Nwanzu[2]

[1]Department of Mathematics and Statistics Federal University Wukari, Taraba State, Nigeria.
[2]Department of Statistics, University of Ibadan, Oyo State, Nigeria.

Accepted 27 February, 2013

Estimation of population parameters using the generalized moment estimators under probability proportional to size sampling scheme requires that the specification parameter, k defining these moments differs from one population to the other due to varying statistical properties of the study and measure of size variables. In this study, the approximate value of the specification parameter that minimizes the anticipated mean squared error is derived and is recommended for use in determining the best estimator for target populations in probability proportional to size (PPS) sampling especially, when the correlation between the study and auxiliary (measure of size) variables is being exploited. Empirical results also confirm that the specification parameter is a useful guide in defining appropriate estimator in PPS multi-character surveys.

Key words: Moment, estimation, specification parameter, transformation, correlation.

INTRODUCTION

Survey statisticians have found probability proportional to size (PPS) sampling scheme useful for selecting units from the population as well as estimating parameters of interest especially when it is clear that the survey is large in size and involves multiple characteristics. Studies on inferences in finite population sampling including the works of Godambe (1955), Basu (1971), and Chaudhuri (2010) have postulated the non-existence of an unbiased estimator of population characteristics with the uniformly least value of its variance. With this development, lots of alternative estimators have been suggested in PPS sampling scheme following the pioneering work by Hansen and Hurwitz (1943). Rao (1966) suggested an alternative estimator in PPS sampling by assuming that the correlation between study variables and measure of size variable is zero. Pathak (1966) proved this theory correct while Bansal and Singh (1985) argued that population correlation can never be zero and provided a non-linear transformation in the selection probabilities. Amahia et al. (1989) developed a transformation that is linear in p_i and possesses the properties of arithmetic mean. This development brought about numerous contributions including the works of Grewal et al. (1997) and Ekaette (2008) involving the linear transformations of the selection probabilities and, Singh et al. (2004) involving non-linear transformation. Recently, Ikughur and Amahia (2011) developed a generalized transformation for a class of alternative linear estimators in PPS sampling scheme within which optimum estimator for any target population is located. An interesting feature of these estimators is that it is defined by the k^{th} moment in correlation coefficient as are related with the statistical properties of target population namely, coefficients of variation, determination, skewness and kurtosis.

In this study, the estimators defined by moments in the linear regression model is considered in order to determine the approximate values of *k* that minimizes the variance of the estimator for target populations.

*Corresponding author. E-mail: atsua2004@yahoo.com.

METHODOLOGY

Consider the homogenous linear estimator (HLE) of the form

$$\hat{t}_g = \sum_{i\in\Omega} b_{si}\, I_{si}\, y_i \text{ , i= 1,2,3, …, N} \quad (1)$$

Where $I_{si}=\begin{cases}0, & if\ i\notin S\\ 1, & if\ i\in S\end{cases}$ and b_{si} are weights not depending on y_i's but on the sample design. Let $\hat{t}_{g,c}$ be the estimators of the population total defined by the generalized transformation *g* under the c^{th} moments, then under probability proportional to size with replacement (PPSWR) sampling, $b_{si} = 1/(np_{i,g}^*)$ so that our estimator of population total becomes:

$$\hat{t}_{g,c} = \frac{1}{n}\sum_{i\in\Omega}\frac{y_i}{p_{i,g}^*} \quad (2)$$

Where

$$p_{i,g,c}^* = \frac{1-g(k)}{N} + g(k)p_i, \text{ with } g(k) = \rho^k,\ k = 1,2,3,4 \quad (3)$$

$0<\rho<1$, $k>0$ so that Equation (2) is the generalized PPS estimators. Ikughur and Amahia (2011) have defined $p_{i,g}^*$ as the generalized transformation for selection probabilities that provide the specification parameter *k*, which is also a pivot element in defining estimators in the linear class. To demonstrate this fact, the generalized estimator can be classified into special cases as follows:

(i) When ρ = 0, $p_{i,g}^* = \frac{1-g(k)}{N} + g(k)p_i$, which is the Rao'(1966) estimator with $p_{i,g}^* = \frac{1}{N}$ irrespective of *k*;

(ii) When ρ = 1, $p_{i,g}^* = \frac{1-g(k)}{N} + g(k)p_i$, which is the Hansen-Hurwitz estimator with $p_{i,g}^* = p_i$;

(iii) When 0 < ρ < 1, $p_{i,g}^* = \frac{1-g(k)}{N} + g(k)p_i$ the moment estimators are defined. Specifically, when *k* =1, Amahia et al. (1985) the estimators is realized;

(iv) In (iii), when Ikughur and Amahia (2011) have postulated the occurrence of the values of *k* = 1, 2, 3 and 4 when 0 < ρ < 25, 25 < ρ < 0.5, 0.5 < ρ < 0.75 and 0.75 < ρ < 1 respectively.

Moment in correlation coefficient

Dodge and Rousson (2000, 2001) showed that, in the context of linear models, the response variable will always have less skew than the explanatory variable and this also applies to the kurtosis of the two variables. Thus, under standard assumptions for linear models, the response variable is a linear combination of an explanatory variable that need not be normally distributed and a normally distributed error. If the explanatory variable has a non-normal distribution, and the error term is normally distributed, then the response variable is a linear convolution of these two distributions that must have a distribution closer to normality than the explanatory variable. We draw inspiration from the works of Dodge and Rousson (2000, 2001) and other workers such as Rovine and von Eye (1997), Rodgers and Nicewander (1988) among others that have established theoretical relationship between correlation coefficient and other statistical properties that are related by the linear model to make the proposition below.

Proposition 1

Consider the linear model

$$y = \beta x + \varepsilon \quad (4)$$

Where y is the response variable, x is the explanatory variable, β is the slope parameter and ε is the error term, then, the expected value of the c^{th} standardized moment of the study variable is given by

$$E\left(\frac{y-\mu_y}{\sigma_y}\right)^k = E\left[\rho\left(\frac{x-\mu_x}{\sigma_x}\right) + R_{\sigma_{\varepsilon,y}}\left(\frac{\varepsilon-\mu_\varepsilon}{\sigma_\varepsilon}\right)\right]^k, k = 1,2,3,4 \quad (5)$$

Proof

From Equation (4), we have

$$y - \mu_y = \beta(x-\mu_x) + (\varepsilon-\mu_\varepsilon) \quad (6)$$

Standardizing Equation (6), we obtain

$$\frac{y-\mu_y}{\sigma_y} = \frac{\beta(x-\mu_x)}{\sigma_y} + \frac{(\varepsilon-\mu_\varepsilon)}{\sigma_y}$$
$$= \frac{\rho\sigma_y(x-\mu_x)}{\sigma_x\sigma_y} + \frac{\sigma_\varepsilon(\varepsilon-\mu_\varepsilon)}{\sigma_\varepsilon\sigma_y}$$
$$= \frac{\rho(x-\mu_x)}{\sigma_x} + \left(\frac{\sigma_\varepsilon}{\sigma_y}\right)\frac{(\varepsilon-\mu_\varepsilon)}{\sigma_\varepsilon}$$

The k^{th} moment of the standardized variable *y* is:

$$\left(\frac{y-\mu_y}{\sigma_y}\right)^c = \left[\frac{\rho(x-\mu_x)}{\sigma_x} + \left(\frac{\sigma_\varepsilon}{\sigma_y}\right)\frac{(\varepsilon-\mu_\varepsilon)}{\sigma_\varepsilon}\right]^c$$

whose expectation is

$$E\left(\frac{y-\mu_y}{\sigma_y}\right)^k = E\left[\rho\left(\frac{x-\mu_x}{\sigma_x}\right) + R_{\sigma_{\varepsilon,y}}\left(\frac{\varepsilon-\mu_\varepsilon}{\sigma_\varepsilon}\right)\right]^k$$

where $R_{\sigma_{\varepsilon,y}} = \left(\frac{\sigma_\varepsilon}{\sigma_y}\right)$ is the ratio of the standard deviation of the error term to the standard deviation of y.

Expression (5) is the generalized expression for expectation of the k^{th} standardized moment. Specifically, when *k* = 1, we have

$$\Phi_{y,1} = \rho^1\Phi_{x,1} + R_{\sigma_{\varepsilon,y}}\Phi_{\varepsilon,1} \quad (7)$$

Where

$$\Phi_{y,1} = E\left(\frac{y-\mu_y}{\sigma_y}\right)^1, \quad \Phi_{x,1} = E\left(\frac{x-\mu_x}{\sigma_x}\right)^1 \text{ and } \quad \Phi_{\varepsilon,1} = E\left(\frac{\varepsilon-\mu_\varepsilon}{\sigma_\varepsilon}\right)^1$$

respectively, so that:

$$\rho^1 = \frac{\Phi_{y,1}}{\Phi_{x,1}} = 0 \quad (8)$$

When $k = 2$, then

$$\Phi_{y,2} = \rho^2 \Phi_{x,2} + R^2_{\sigma_{\varepsilon,y}} \Phi_{\varepsilon,2} \quad (9)$$

But $\Phi_{x,2} = 1;\ \ \Phi_{\varepsilon,2} = 1;\ \text{and also}\ \Phi_{y,2} = 1$.

Therefore,

$$1 = \rho^2 + R^2_{\sigma_{\varepsilon,y}} \quad (10)$$

Under linear framework, $R^2_{\sigma_{\varepsilon,y}} < 1$ always. At this point, two scenarios can be identified, namely:

(i) When $R^2_{\sigma_{\varepsilon,y}}$ is negligible. Here, $\rho^2 \to 1.$

(ii) When $R^2_{\sigma_{\varepsilon,y}}$ is a quantity in [0,1] and $\rho^2 \text{does not tend to } 1.$

when $k = 3$, then

$$\Phi_{y,3} = \rho^3 \Phi_{x,3} + 3\rho^2 R_{\sigma_{\varepsilon,y}} \Phi_{\varepsilon,1} \Phi_{x,2} + 3\rho R^2_{\sigma_{\varepsilon,y}} \Phi_{\varepsilon,2} \Phi_{x,1} + R^3_{\sigma_{\varepsilon,y}} \Phi_{\varepsilon,3} \quad (11)$$

So that

$$\gamma_y = \rho^3 \gamma_x + R^3_{\sigma_{\varepsilon,y}} \gamma_\varepsilon \quad (12)$$

where $\gamma_y = \Phi_{y,3};\ \gamma_x = \Phi_{x,3}\ \text{and}\ \gamma_\varepsilon = \Phi_{\varepsilon,3}$ are the skewness coefficients of y, x and ε respectively. If $R^3_{\sigma_{\varepsilon,y}}$ and hence $R^3_{\sigma_{\varepsilon,y}} \gamma_\varepsilon$ are negligible, then

$$\rho^3 = \frac{\gamma_y}{\gamma_x},\ \gamma_x \neq 0. \quad (13)$$

$\Rightarrow \gamma_y < \gamma_x$ satisfying $0 < \rho^3 < 1$.

Remark 1

Expression (13) expresses the third power of correlation coefficient as the ratio of the skewness coefficient of y and x. For $k = 4$,

$$\Phi_{y,4} = \rho^4 \Phi_{x,4} + 6\rho^2 R^2_{\sigma_{\varepsilon,y}} \Phi_{\varepsilon,2} \Phi_{x,2} + R^4_{\sigma_{\varepsilon,y}} \Phi_{\varepsilon,4} \quad (14)$$

So that

$$\rho^4 = \frac{K_y}{K_x};\ K_x \neq 0 \quad (15)$$

Expression (15) represents the forth power of correlation coefficient as the ratio of the kurtosis of the response variable and the kurtosis of the explanatory variable. This can be interpreted as the percentage of kurtosis which is presented by linear model.

Certainly, $\rho^4 = \frac{K_y}{K_x} < 1$ is expected to hold true. Similarly, under Equation (4)

$$E(y) = \beta E(x) + E(\varepsilon)$$

So that

$$\rho^1 = \frac{\mu_y}{\mu_x} \frac{\sigma_x}{\sigma_y} = \frac{CV_x}{CV_y} \quad (16)$$

By this proposition, moments in correlation coefficient have been linked with statistical properties namely, coefficients of variation, determination, skewness and kurtosis of target populations as follows:

(i) $\rho^1 = \frac{CV_x}{CV_y}$; with $CV_y \neq 0;$

(ii) $\rho^2 = 1 - R^2_{\sigma_{\varepsilon,y}}$; $R^2_{\sigma_{\varepsilon,y}}$ does not tend to zero.

(iii) $\rho^3 = \frac{\gamma_y}{\gamma_x}$ <1; $\gamma_x \neq 0$ and

(iv) $\rho^4 = \frac{K_y}{K_x}$ <1; $K_x \neq 0$.

However, an alternative approach to the determination of the specification parameter is sought at this point. To achieve this objective, the expected variance formula for which an approximate expression is realized under the super-population model is utilized.

Super-population model inference

Stochastic model for Y is called super-population model (Deming and Stephan, 1941; Korn and Graubard, 1998) and have been used extensively to evaluate design and estimators (Cochran, 1946; Cassel et al., 1997) among others. Here, we consider the super-population model defined by

$$y = \beta p_i + \varepsilon \quad (17)$$

With $E(\varepsilon/p_i) = 0,\ Cov(\varepsilon_i \varepsilon_j) = 0\ \text{and}\ E(\varepsilon_i^2) = a p_i^g$

Under the PPS sampling design, the mean square error (MSE) is defined as:

$$= \frac{1}{n}\left[\sum_{i\epsilon\Omega} \frac{I_{si} y_i^2 p_i}{p_{i,g}^{*2}} - \left(\sum_{i\epsilon\Omega} \frac{I_i y_i p_i}{p_i^*}\right)^2\right] + \left[\sum_{i\epsilon\Omega} \left(\frac{I_{si} p_i}{p_{i,g}^*} - 1\right) y_i\right]^2 \quad (18)$$

whose bias is

$$B_p(\hat{\tau}_{g,c}) = \sum_{i\epsilon\Omega} \left(\frac{I_{si} p_i}{p_{i,g}^*} - 1\right) y_i \quad (19)$$

Theorem 1

The expected MSE of the proposed estimator under super-population model is:

$$\xi V_e(\hat{t}_{g,c}) = \frac{1}{n}\left[\sum_{i\epsilon\Omega}\frac{I_{si}\xi(y_i^2)p_i}{p_{i,g}^{*2}} - \left(\sum_{i\epsilon\Omega}\frac{I_i\xi(y_i)p_i}{p_{i,g}^{*}}\right)^2\right] + \left[\sum_{i\epsilon\Omega}\left(\frac{I_{si}p_i}{p_{i,g}^{*}} - 1\right)\xi(y_i)\right]^2 \quad (20)$$

Proof

It follows by substituting Equations (18) and (19) in the expression of MSE defined by

$$MSE(\hat{t}_{g,c}) = V_p(\hat{t}_{g,c}) + B^2(\hat{t}_{g,c})$$

Considering the design based MSE above, the expected variance of the conventional estimator is

$$\xi V_p(\hat{t}_c) = \frac{1}{n}\left[\sum_{i\epsilon\Omega}\frac{I_{si}\xi(y_i^2)}{p_i} - (\sum_{i=1}^{N}\xi(y_i))^2\right] = \frac{a}{n}\left[\sum_{i\epsilon\Omega}p_i^{g-1} - \sum_{i=1}^{N}p_i^{g}\right] \quad (21)$$

Similarly the expected variance of the generalized estimator is

$$\xi V\ (\hat{t}_{g,c}) = \frac{1}{n}\left[\sum_{i\epsilon\Omega}\frac{I_{si}\xi(y_i^2)p_i}{p_{i,g}^{*2}} - \left(\sum_{i\epsilon\Omega}\frac{I_i\xi(y_i)p_i}{p_i^{*}}\right)^2\right] = \frac{a}{n}\left[\sum_{i\epsilon\Omega}\frac{p_i^{g+1}(1-p_i)}{p_{i,g}^{*2}}\right] + \frac{\beta^2}{n}\left[\sum_{i\epsilon\Omega}\frac{p_i^{3}}{p_{i,g}^{*2}} - (\sum_{i=1}^{N}\frac{p_i^{2}}{p_i^{*}})^2\right] \quad (22)$$

Now, comparing the two variances, we have $n[\xi V_p(\hat{t}_{g,c}) - \xi V_p(\hat{t}_c) = n\nabla]$ so that

$$n\nabla = a\sum_{i\epsilon\Omega}\frac{p_i^{g+1}(1-p_i)}{p_{i,g}^{*2}} + \beta^2\sum_{i\epsilon\Omega}\frac{p_i^{3}}{p_{i,g}^{*2}} - (\sum_{i=1}^{N}\frac{p_i^{2}}{p_i^{*}})^2 - (\sum_{i\epsilon\Omega}p_i^{g-1} - \sum_{i=1}^{N}p_i^{g})$$

$$= a\sum_{i\epsilon\Omega}\frac{p_i^{g+1}(1-p_i)}{p_{i,g}^{*2}} - \sum_{i\epsilon\Omega}p_i^{g}(\frac{1-p_i}{p_i}) + \beta^2 V\left(\frac{p_i}{p_i^{*}}\right) \quad (23)$$

Now, let $C = \sum_{i\epsilon\Omega}\frac{p_i^{g+1}(1-p_i)}{p_{i,g}^{*2}} - \sum_{i\epsilon\Omega}p_i^{g}(\frac{1-p_i}{p_i})$

$$= a\sum_{i\epsilon\Omega}\frac{p_i^{g-1}(1-p_i)}{p_{i,g}^{*2}}(p_i^2 - p_i^{*2}) \quad (24)$$

Satisfying $n\nabla = aC+\beta^2 D$ (25)

empirically, when ρ=0, $D = \mathrm{Var}(p_i/p_i^*) > 0$ (26)

and as ρ > 0, $D = \mathrm{Var}(p_i/p_i^*) \rightarrow 0$ (27)

In most real life scenario, ρ≠0 always, we consider Equation (27) as most ideal for surveys and hence, inference based on *aC* will be sufficient.

Now, let

$$C = \sum_{i=1}^{n} b_i^* c_i^* = \sum_{i\epsilon\Omega}\frac{p_i^{g-1}(1-p_i)}{p_{i,g}^{*2}}(p_i^2 - p_i^{*2}) \quad (28)$$

Where

$$b_i^* = \frac{p_i^{g-1}(1-p_i)}{p_{i,g}^{*2}} \quad (29)$$

And

$$c_i^* = (p_i^2 - p_i^{*2}) \quad (30)$$

Then, we can as well observe that

$$\sum_{I=1}^{n} c_i^* < 0,\ if\ 0 < \rho < 1,$$

or

$$\sum_{i=1}^{n} c_i^* = 0 \text{ if } \rho = 0 \text{ or } \rho = 1.$$

negligible as $\xi B^2(\hat{t}_{g,c}) = \nabla \rightarrow 0$

Proof

Considering the anticipated bias from the model,

$$\xi B(\hat{t}_{g,c}) = \sum_{i\epsilon\Omega}\left(\frac{I_{si}p_i}{p_{i,g}^{*}} - 1\right)\xi(y_i)$$

$$= \beta\sum_{i\epsilon\Omega}\left(\frac{I_{si}p_i}{p_{i,g}^{*}} - 1\right)p_i$$

when $\frac{p_i}{p_i^{*}} = 1$ then $\beta\sum_{i\epsilon\Omega}\left(\frac{I_{si}p_i}{p_{i,g}^{*}} - 1\right)p_i = 0$ and $B(\hat{t}_{g,c}) = 0$

when $\frac{p_i}{p_i^{*}} < 1$ then $\beta\sum_{i\epsilon\Omega}\left(\frac{I_{si}p_i}{p_{i,g}^{*}} - 1\right)p_i < 1$ and hence, $B(\hat{t}_{g,c}) \rightarrow \nabla < 1$ especially when $\beta \rightarrow 1$.

Also, when $\frac{p_i}{p_i^{*}} > 1$ then $\beta\sum_{i\epsilon\Omega}\left(\frac{I_{si}p_i}{p_{i,g}^{*}} - 1\right)p_i < 1$ and $B(\hat{t}_{g,c}) \rightarrow \nabla < 1$ especially when $\beta \rightarrow 1$. Since $B(\hat{t}_{g,c}) = 0$ when $\frac{p_i}{p_i^{*}} = 1$ which is a necessary condition for unbiasness, we can conveniently state that in the case of a biased estimator the condition becomes $0 < B(\hat{t}_{g,c}) < \nabla$.

Alternatively, by Cauchy-Schwarz inequalities,

$$\xi B^2(\hat{t}_{g,c}) = \left[\beta\sum_{i\epsilon\Omega}\left(\frac{I_{si}p_i}{p_{i,g}^{*}} - 1\right)p_i\right]^2 = \beta^2\left[\sum_{i\epsilon\Omega}\left(\frac{I_{si}p_i}{p_{i,g}^{*}} - 1\right)p_i\right]^2 \leq \beta^2\left[\sum_{i\epsilon\Omega}\left(\frac{p_i}{p_i^{*}} - 1\right)^2\sum_{i\epsilon\Omega}p_i^2\right]$$

But $\sum_{i\epsilon\Omega}\left(\frac{p_i}{p_i^{*}} - 1\right)^2\sum_{i\epsilon\Omega}p_i^2 = \nabla < 1$ so that $0 < \left[\sum_{i\epsilon\Omega}\left(\frac{p_i}{p_i^{*}} - 1\right)^2\sum_{i\epsilon\Omega}p_i^2\right] < 1$ always.

Therefore, $\xi B^2(\hat{t}_{g,c}) = \nabla \rightarrow 0.$

Table 1. Specification parameters k for alternative linear estimators for four study populations at varying g.

Population	g = 0		g = 1		g = 2	
	Min p_i	Max p_i	Min p_i	Max p_i	Min p_i	Max p_i
I	0	2	0	2	0	1
II	0	2	0	4	0	2
III	0	3	0	3	0	2
V	0	13(4)	0	6(4)	0	8(4)

*Values in parenthesis are ceiling values that are used instead of the original values.

Thus, under super-population model, the expected bias per unit is very negligible especially when $\beta \leq 1$ as such, inference based on the expected variance would be sufficient.

Determination of approximate values of *k*

Studies have shown that the value of *g* useful in estimation range

Theorem 2

Under super-population model, the expected bias per unit is very between 0 and 2 inclusive. Amahia (1987) have shown that the value of *g* is given by

$$g > \frac{2\rho p_i}{p^*_{i,g}} + \frac{1}{1-p_i} - \frac{(1+\rho)p_i}{p^*_{i,g}+p_i} \quad (31)$$

Now, from Equation (31) we define

$$b_i^* = \frac{p_i^{g-1}(1-p_i)}{p_{i,g}^{*2}}$$

Such that $\frac{db_i^*}{dp_i} = \frac{p_i^{*2}\frac{d}{dp_i}\left[(1-p_i)p_i^{g-1}\right] - (1-p_i)p_i^{g-1}\frac{d}{dp_i}p_{i,g}^{*2}}{p_{i,g}^{*4}} = 0$

$$\Rightarrow p_{i,g}^{*2}\left[(g-1)p_i^{g-2} - gp_i^{g-1}\right] - 2\rho^c p_{i,g}^* p_i^{g-1}(1-p_i) = 0$$

$$\left[\left[\frac{1-\rho^c}{N} + \rho^c p_i\right](g-1)p_i^{g-2} - gp_i^{g-1}\right] - 2\rho^c p_i^{g-1}(1-p_i) = 0$$

$$\Rightarrow Ap_i^{g-2} - A\rho^c p_i^{g-2} - Bp_i^{g-1} + B\rho^c p_i^{g-1} + AN\rho^c p_i^{g-1} - g\rho^c p_i^g - 2\rho^c p_i^{g-1} + 2\rho^c p_i^g = 0$$

$$\Rightarrow \Phi_{1,p,g} = \Phi_{2,p,g}\rho^c$$

$$\rho^c = \frac{\Phi_{1,p,g}}{\Phi_{2,p,g}} = |\gamma|$$

Therefore,

$$k \cong \left|\frac{\log(\gamma)}{\log(\rho)}\right| \quad (32)$$

Where

$$A = \frac{g-1}{N},\ B = \frac{g}{N};$$

$$\Phi_{1,p,g} = Ap_i^{g-2} - Bp_i^{g-1}$$

And

$$\Phi_{2,p,g} = Ap_i^{g-2} - Bp_i^{g-1} - ANp_i^{g-1} + gp_i^g + 2p_i^{g-1} - 2p_i^g$$

Under the k^{th} standardized moment of the study variable utilized in this study, the limiting value of *k*, the specification parameter is 4. Thus, there are N values of *k* thereby giving rise to a range of values of *k*.

Empirical illustration

For illustration, four study populations namely, populations I, II, III and IV with ρ = 0.162, ρ = 0.395, ρ = 0.51 and ρ = 0.91, respectively describing four moment conditions of “weak correlation”, “moderate correlation”, “ high correlation” and “very high correlation” are used in the Expression (29) to obtain the approximate values of *k*. The results are shown on Table 1. Since the distributions consist of N values of p_i, N values of the specification parameters *k* required. For convenience, our interest is reduced to obtaining the lower and upper values of *k* determined by Min p_i and Max p_i. From Table 1, it is clear that the true value of *k* is [0,2], [0,2], [0,3] and [0,13] for populations I, II, III and IV respectively when g=0. However, when g =1, c is a value in [0,2], [0,4], [0,3] and [0,6] for Populations I, II, III and IV, respectively. When g=2, c is a value in [0,1], [0,2], [0,2] and [0,8] for populations I, II, III and IV respectively. If the integer values are considered (since we are considering the integer moments only) noting that $k \neq 0$, then the values of k determined by Max p_i will be sufficient. Again, if the ceiling values of *k* is 4 (as described by the coefficient of kurtosis), then we assume that k = 4 is adequate when the estimated value of *k* is greater than 4.

Conclusion

In view of the non-existence of a uniformly most efficient

estimator theory, this study brings a new dimension in defining a linear estimator in PPS sampling design. The estimators in this class are only specified when the statistical properties of the study populations in relationship with correlation coefficients are assumed known. The prior knowledge of selection probabilities is also assumed which helps the survey statistician to have a prior knowledge of the range value of the specification parameter before embarking on estimation of population characteristics. Specifically, the values of k = 1, 2, 2 and 4 are adequate to define estimators in Equation (2) for Populations I, II, III and IV, respectively.

REFERENCES

Amahia GN (1987). Some Sampling Strategies under Super-population Models. Unpublished PhD Thesis. University of Ibadan, Nigeria.

Amahia GN, Chaubey YP, Rao TJ (1989). Efficiency of a new PPS sampling for multiple characteristics. J. Stat. Plan. Inference 21:75-84.

Bansal ML, Singh R (1989). An alternative estimator for multiple characteristics in PPS sampling. J. Stat. Plan. Inference 21:75-84.

Basu D (1971). An essay on the logical foundation of survey sampling I" Foundation of statistical inference. Holt, Rinehard and Winston. Edited by Godambe and Sprott.

Cassel, CM, Sarndal C, Wretman JH (1977). Foundation of Inference in Survey Sampling. Wiley Interscience Publication, New York. pp. 80-81.

Chaudhuri A (2010). Essentials of Survey Sampling. PHI Learning Private Limited, New Delhi. pp. 12-14.

Cochran WG (1946). Relative accuracy of systematic and stratified random samples for a certain class of population. Ann. Math. Stat. 17:164-177.

Deming WE, Stephan FF (1941). On the interpretation of censuses as samples. J. Am. Stat. Assoc. 36:45-49.

Dodge Y, Rousson V (2000). Direction of Dependence in a regression line. Commun. Stat. Theory Methods 29(9-10):1945-1955.

Dodge Y, Rousson V (2001). On asymmetric properties of the correlation coefficient in the regression setting. Am. Stat. 55:51-54.

Ekaette IE (2008). A class of alternative estimators for Multi-characteristics in PPS sampling Scheme. Unpublished Ph.D thesis, Univeristy of Ibadan. Nigeria.

Godambe VP (1955). A unified theory of sampling from finite population. J. Roy. Stat. Soc. B. 17:269-278.

Grewal IS, Bansal ML, Singh (1997). An alternative estimator for multiple characteristics using randomized response technique in PPS sampling. Aligarh J. Stat. 19:51-65.

Hansen MH, Hurwitz WN (1943). On the theory of sampling from a finite population. Ann. Math. Stat. 14:333-362.

Ikughur AJ, Amahia GN (2011). A generalized transformation for selection probabilities in unequal probability sampling scheme. J. Sci. Ind. Stud. 9(1):58-62.

Korn EL, Graubard BI (1998). Variance estimation for superpopulation parameters. Statistica Sinica 8:1131-1151.

Pathak PK (1966). An estimator in PPS sampling for multiple characteristics. *Sankhya, A.* 28(1):35-40.

Rao JNK (1966). Alternative estimators in the PPS sampling for multiple characteristics. Sankhya 28(A):47-60.

Rodgers JL, Nicewander WA (1988). Thirteen ways to look at the correlation coefficient. Am. Stat. 42:59-66.

Rovine MJ, von Eye A (1997). A 14th way to look at a correlation coefficient: Correlation as the proportion of matches. Am. Stat. 51:42-46.

Singh S, Grewal IS, Joarder A (2004). General class of estimators in multi-character surveys. Stat. Papers 45:571-582.