

Comparison of Some Sample Selection Procedures In Probability Proportional to Size

Omotola O. Dawodu & Adedayo A. Adewara,

Department of Statistics,
University of Ilorin,
Ilorin,
Nigeria

ABSTRACT

The importance of probability proportional to size sampling cannot be over emphasized. In recent times, Statisticians often work on complex survey data which involve weighting, as a result of the unequal probability nature of units. There is need to work on design weights, which is defined as the inverse of the probability of inclusion of single unit in the sample. Two simulated data sets are generated using Rin order to obtain the probability of inclusion of single unit and also the joint probability of inclusion of units in the sample. Auxiliary variables and variables of interest were generated in such a way that there exists reasonable correlation between both. A procedure for selecting samples with probability proportional to size nature without replacement was proposed. This procedure could be used with Horvitz and Thompson estimator of variance. The variance estimates are obtained using the modified sample selection procedure and an existing procedure, and results obtained. For the two data sets, it is observed that the modified procedure is more efficient than the existing one.

Keywords: Inclusion probability, probability sampling, sampling scheme, unequal probability sampling, sampling without replacement.

INTRODUCTION

The role of the Horvitz and Thompson estimator of the population total in probability proportional to size sampling without replacement (*ppsswor*) cannot be over emphasized. [3] gave the estimator of the population total in *ppsswor* as:

$$\hat{Y}_{HT} = \sum_{i=1}^N \frac{Y_i}{\pi_i} \quad (1)$$

Correspondence: Omotola O. Dawodu, University of Ilorin, Ilorin, Nigeria,
E-mail: omotola_dawodu@yahoo.com

and the unbiased variance estimator as:

$$\hat{V}(\hat{Y}_{HT}) = \sum_{i=1}^n \left(\frac{1-\pi_i}{\pi_i^2} \right) y_i^2 + \sum_{i=1}^n \sum_{j \neq i} \left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}} \right) \left(\frac{y_i y_j}{\pi_i \pi_j} \right) \quad (2)$$

which may give negative results.

[4] and [7] independently gave the expression for the unbiased variance of Horvitz and Thompson estimator as follows:

$$\hat{V}_{SYG}(\hat{Y}_{HT}) = \sum_{i=1}^n \sum_{j \neq i}^n \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i y_j}{\pi_i \pi_j} \right)^2 \quad (3)$$

They showed that their variance estimator produces positive results for most sampling schemes.

Y_i represents the variable of interest. π_i is the probability of inclusion of single unit i in the sample, while π_{ij} is the joint probability of inclusion of both units i and j in the sample.

There are many sample selection procedures that survey statisticians have proposed that could be used with (3) above, some of which are: [1], [2], [5], and [6].

1. Methods

1.1 Alodat Sample Selection Procedure

[1] proposed a sample selection procedure in *ppssworr* as follows:

- Select the first unit with probability proportional to $\frac{P_i}{2-P_i}$
- Select second unit with probability $\frac{P_i}{1-P_j}$.

Therefore,

$$\pi_i = \frac{P_i}{B} \left[\frac{1-2P_i}{(2-P_i)(1-P_i)} + \sum_{j=1}^N \frac{P_j}{(2-P_j)(1-P_j)} \right] \quad (4)$$

$$\text{where } B = \sum_{i=1}^N \frac{P_i}{2-P_i}$$

$$\pi_{ij} = \frac{P_i P_j}{B} \left[\frac{1}{(2-P_i)(1-P_i)} + \frac{1}{(2-P_j)(1-P_j)} \right] \quad (5)$$

This procedure is for selecting sample of size two.

1.2 Basit and Shahbaz Sample Selection Procedure (BSP)

[2] proposed a sample selection procedure in *ppsswor* as follows:

- Select first unit with probability proportional to $\frac{P_i^\alpha(1-P_i^\beta)}{(1-2P_i^\beta)}$
- Select a random sample of size $(n-1)$ from the remaining $(N-1)$ units.

Therefore

$$\pi_i = \frac{1}{d} \left[\frac{P_i^\alpha(1-P_i^\beta)}{(1-2P_i^\beta)} \left(\frac{N-n}{N-1} \right) + \frac{(n-1)}{N-1} d \right] \quad (6)$$

$$\text{where } d = \sum_{i=1}^N \frac{P_i^\alpha(1-P_i^\beta)}{(1-2P_i^\beta)}$$

$$\pi_{ij} = \frac{(n-1)(N-n)}{d(N-1)(N-2)} \left[\frac{P_i^\alpha(1-P_i^\beta)}{(1-2P_i^\beta)} + \frac{P_j^\alpha(1-P_j^\beta)}{(1-2P_j^\beta)} \right] + \frac{(n-1)(n-2)}{(N-1)(N-2)} \quad (7)$$

where α and β are constants.

1.3 Modified Procedure (MP)

Assume that we have total number of units N and it is desired to select a sample of size n from N , where $n < N$. A sample selection procedure that uses the ideas of [1] and [2] is proposed. The modified procedure (MP) can be used to select sample of any size in *ppsswor*.

The procedure is as follows:

- Select the first unit with probability proportional to $\frac{P_i^\alpha}{2-P_i^\beta}$
- Select a random sample of size $n-1$ from the remaining $N-1$ units, for given α and β .
- Note that α and β are constant values and are expressed in powers.

Therefore,

$$\pi_i = \frac{\frac{P_i^\alpha}{2-P_i^\beta}}{\sum_{i=1}^N \frac{P_i^\alpha}{2-P_i^\beta}} + \frac{n-1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^N \frac{\frac{P_j^\alpha}{2-P_j^\beta}}{\sum_{j=1}^N \frac{P_j^\alpha}{2-P_j^\beta}}$$

$$\text{Let } R = \sum_{i=1}^N \frac{P_i^\alpha}{2 - P_i^\beta}$$

$$\pi_i = \frac{1}{R} \left[\frac{P_i^\alpha}{2 - P_i^\beta} + \frac{n-1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^N \frac{P_j^\alpha}{2 - P_j^\beta} \right]$$

$$\text{But } \sum_{\substack{j=1 \\ j \neq i}}^N \frac{P_j^\alpha}{2 - P_j^\beta} = \sum_{j=1}^N \frac{P_j^\alpha}{2 - P_j^\beta} - \frac{P_i^\alpha}{2 - P_i^\beta}$$

$$\pi_i = \frac{1}{R} \frac{N-n}{N-1} \frac{P_i^\alpha}{2 - P_i^\beta} + \frac{n-1}{N-1} \quad (8)$$

$$\pi_{ij} = \frac{(n-1)(N-n)}{R(N-1)(N-2)} \left[\frac{P_i^\alpha}{2 - P_i^\beta} + \frac{P_j^\alpha}{2 - P_j^\beta} \right] + \frac{(n-1)(n-2)}{(N-1)(N-2)} \quad (9)$$

The selection probabilities P_i s are calculated from the auxiliary variable X_i s, for $i = 1, 2, \dots, N$.

$P_i = \frac{X_i}{X}$ where X_i is the number of items contained in unit i of the auxiliary variable, X is the total number of items contained in the auxiliary variable.

The P_i s are used to calculate the probability of inclusion of unit (π_i) in the sample and also the joint probability of inclusion (π_{ij}) of both units i and j in the sample. The variance estimates are calculated using (3) above.

2. Some Results for the Modified Procedure

Result 1:

$$\sum_{i=1}^N \pi_i = n$$

Result 2:

$$\sum_{\substack{j=1 \\ j \neq i}}^N \pi_{ij} = (n-1)\pi_i$$

Result 3:

$$\sum_{\substack{j=1 \\ j \neq i}}^N \sum_{i=1}^N \pi_{ij} = n(n-1)$$

Result 4: The Sen – Yates – Grundy variance estimator is always non – negative under this modified sample selection procedure.

3. Presentation and Analysis of Data

We generated auxiliary variables X_i and the variables of interest Y_i , using statistical package *R*, and ensuring that there exists reasonable and positive correlation between them. Hence, simulated Data I with seven units and simulated Data II with nine units were obtained and presented in Tables 1 and 6 respectively. For simplicity, a sample of size six was selected from Data I, while a sample of size eight was selected from Data II. For simplicity, α and β assume integer values, and we worked on the interval -5 to $+5$ for their values in order to see clearly the behaviors of the variance estimates as α and β vary. The variance estimates for specified values of α and β , are approximated to the nearest whole numbers, using the MP and BSP, for the two data sets, and are presented in Tables 5 and 10 respectively. It was observed from Table 5 that the least variance estimates for Data I was obtained when $\alpha = 5$ and $\beta = -5$ for the PP, and $\alpha = 5$ and $\beta = 1$ for the BSP. For Data II, least variance for PP was obtained at $\alpha = 5$ and $\beta = -5$, and $\alpha = 5$ and $\beta = 2$ for BSP, as seen in Table 10.

Units	Auxiliary variable X_i	Variable of interest Y_i
1	201	48
2	218	60
3	230	75
4	220	65
5	237	80
6	240	92
7	250	102

Source: Simulation

Table 1: Distribution of the auxiliary variable and variable of interest (Simulated Data I)

Units	MP ($\alpha = 5, \beta = -5$)	BSP ($\alpha = 5, \beta = 1$)
1	0.83898	0.8451027
2	0.8460511	0.8512921
3	0.8550673	0.857097
4	0.8472673	0.8521688
5	0.8626661	0.8611414
6	0.8665982	0.86304
7	0.88337	0.8701579
Total	6	6

Source: Researchers' Computation

Table 2: Probability of inclusion, π_i (Data I)

Units Sampled	MP ($\alpha = 5, \beta = -5$)	BSP ($\alpha = 5, \beta = 1$)
1, 2	0.6850312	0.6963948
1, 3	0.6940473	0.7021997
1, 4	0.6862474	0.6972715
1, 5	0.7016461	0.7062441
1, 6	0.7055782	0.7081427
1, 7	0.72235	0.7152606
2, 3	0.7011184	0.7083891
2, 4	0.6933185	0.7034609
2, 5	0.7087172	0.7124335
2, 6	0.7126493	0.7143321
2, 7	0.7294211	0.72145
3, 4	0.7023346	0.7092659
3, 5	0.7177333	0.7182385
3, 6	0.7216654	0.7201371
3, 7	0.7384373	0.7272549
4, 5	0.7099334	0.7133102
4, 6	0.7138655	0.7152089
4, 7	0.7306373	0.7223267
5, 6	0.7292642	0.7241815
5, 7	0.746036	0.7312993
6, 7	0.7499681	0.7331979
Total	15	15

Source: Researchers' Computation

Table 3: Joint Probability of inclusion, π_{ij} (Data I)

At these values of α and β stated above, the probabilities of inclusion of single unit, π_i for both data sets are obtained and presented in Tables 2 and 7 respectively. Also, the joint probabilities of inclusion of units in the sample are presented in Tables 3 and 8 respectively. The variance estimates using both procedures for the two data sets are presented in Tables 4 and 9 respectively. Table 11 shows the summary of the least variance estimates obtained using the MP and BSP for the two data sets.

Comparison of Some Sample Selection Procedures In Probability Proportional to Size

Units Sampled	MP ($\alpha = 5, \beta = -5$)	BSP ($\alpha = 5, \beta = 1$)
1, 2	6.79716	6.193011
1, 3	31.27943	29.72315
1, 4	13.63384	12.45956
1, 5	39.77128	39.69422
1, 6	72.94578	74.30393
1, 7	88.22722	102.6572
2, 3	8.976271	8.69371
2, 4	1.140672	1.049416
2, 5	14.20084	14.56765
2, 6	35.79695	37.19587
2, 7	48.85269	58.46588
3, 4	3.810292	3.755415
3, 5	0.7332039	0.8042168
3, 6	9.119637	9.909919
3, 7	17.63317	22.5286
4, 5	7.581205	7.952997
4, 6	24.74576	26.03146
4, 7	36.60806	44.54834
5, 6	4.52875	4.928989
5, 7	11.09356	14.58229
6, 7	1.796104	2.735598
Total	479	523

Source: Researchers' Computation

Table 4: Least variance estimates (Data I)

		α									
		-5		-4		-3		-2		-1	
β		MP	BSP	MP	BSP	MP	BSP	MP	BSP	MP	BSP
	-5	570	603	562	598	553	592	544	586	535	578
	-4	578	603	570	598	562	592	553	586	544	578
	-3	586	603	578	598	570	592	562	586	553	578
	-2	592	603	585	598	578	593	570	586	561	579
	-1	596	603	590	599	583	593	575	586	567	579
	1	603	602	598	597	592	591	585	584	578	577
	2	603	603	598	598	592	592	586	585	578	578
	3	603	603	598	598	592	592	586	586	578	578
	4	603	603	598	598	592	592	586	586	578	578
	5	603	603	598	598	592	592	586	586	578	578

Source: Researchers' Computation

Table 5: Comparison of variances using MP and BSP (Data I)

		α									
		1		2		3		4		5	
β		MP	BSP	MP	BSP	MP	BSP	MP	BSP	MP	BSP
	-5	516	562	506	553	497	544	488	535	479	525
	-4	525	562	516	553	506	544	497	535	488	525
	-3	534	562	525	553	515	544	506	535	497	525
	-2	543	562	534	553	524	544	515	535	505	525
	-1	549	563	540	554	531	545	521	535	512	526
	1	561	560	552	551	543	542	534	532	524	523
	2	562	562	553	553	544	544	534	534	525	525
	3	562	562	553	553	544	544	535	534	525	525
	4	562	562	553	553	544	544	535	535	525	525
	5	562	562	553	553	544	544	535	535	525	525

Source: Researchers' Computation

Table 5 continues

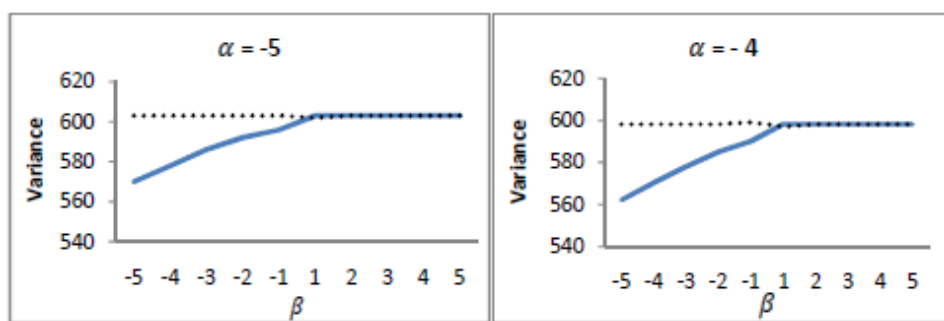


Figure 1(a) —MP BSP

Figure 1(b) —MP BSP

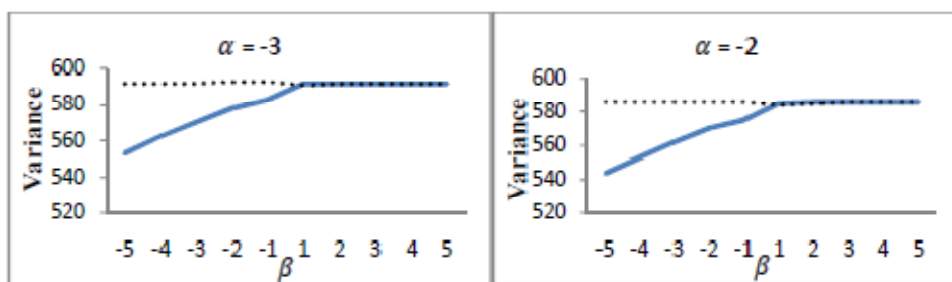


Figure 1(c) —MP BSP

Figure 1(d) —MP BSP

Comparison of Some Sample Selection Procedures In Probability Proportional to Size

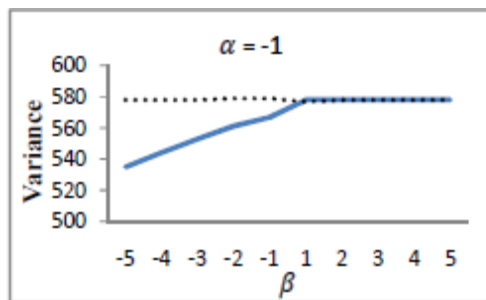


Figure 1(e) —MP BSP

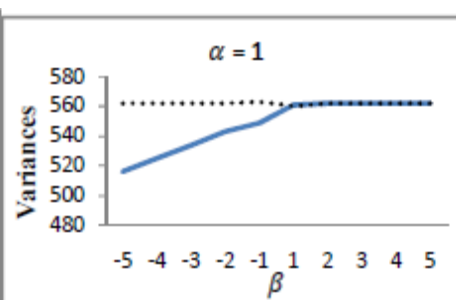


Figure 1(f) —MP BSP

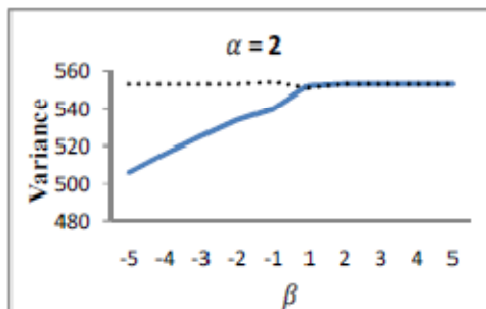


Figure 1(g) —MP BSP

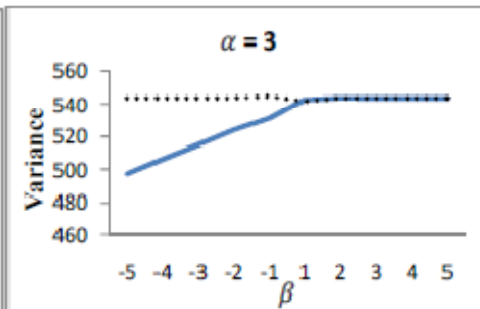


Figure 1(h) —MP BSP

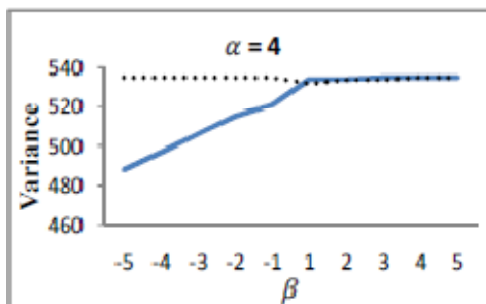


Figure 1(i) —MP BSP

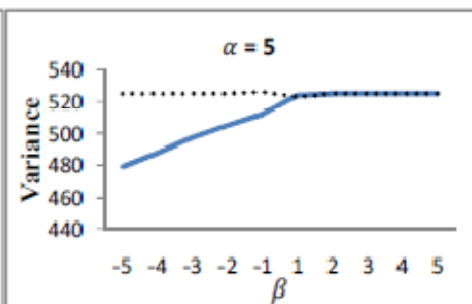


Figure 1(j) —MP BSP

Figures 1: Showing variances using MP and BSP for Data I

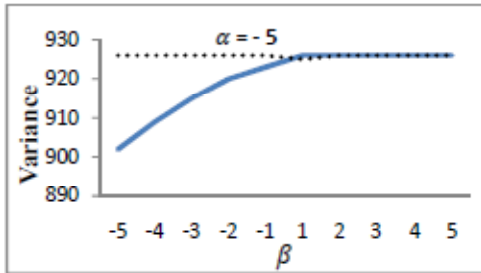


Figure 2 (a) —MP BSP

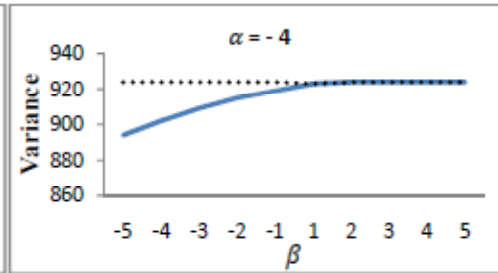


Figure 2 (b) —MP BSP

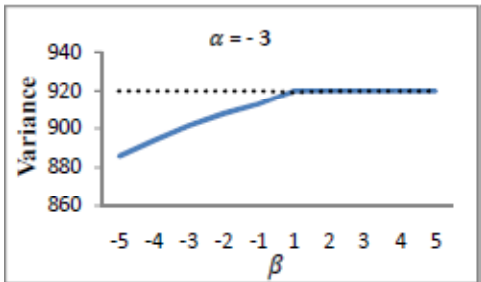


Figure 2 (c) —MP BSP

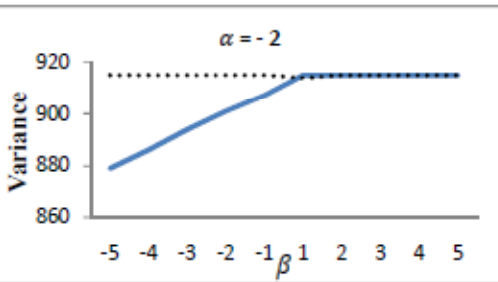


Figure 2 (d) —MP BSP

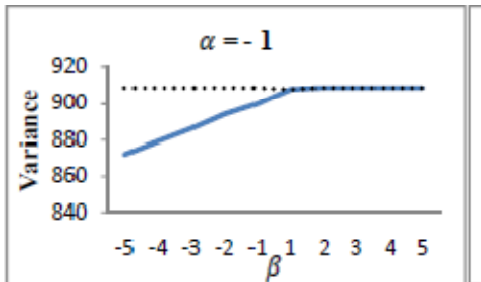


Figure 2 (e) —MP BSP

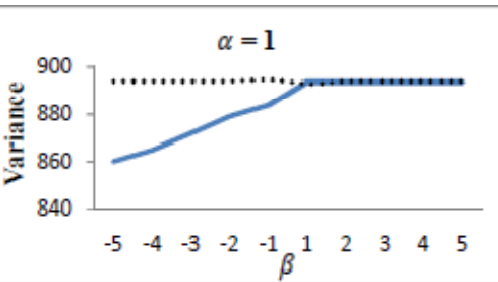


Figure 2 (f) —MP BSP

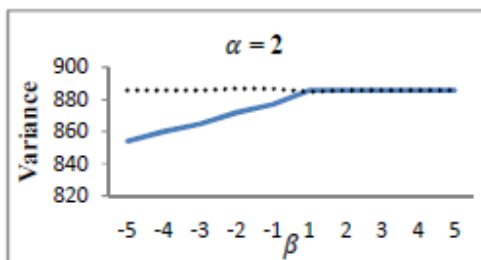


Figure 2 (g) —MP BSP

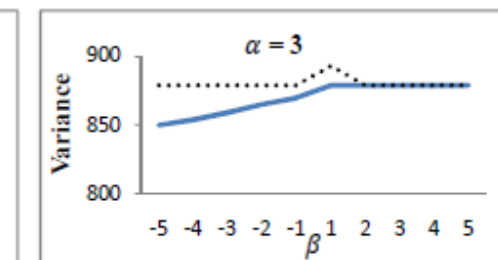


Figure 2 (h) —MP BSP

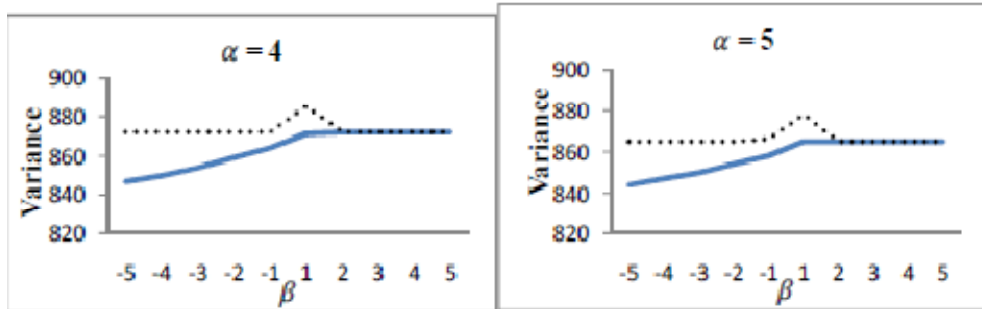


Figure 2 (i) —MP BSP

Figure 2 (j) —MP BSP

Figures 2: Showing variances using MP and BSP for Data II

4. Results and Discussion

It is seen from Tables 5 and 10 and also, Figures 1 and 2 that for all combinations of $\beta = -5$ to -1 , and for all values of α under consideration, the MP produces smaller variance estimates than the BSP. It is observed that for all combinations of $\beta = 1$ to 2 , and for all values of α under consideration, BSP performs better than MP with slight differences. For $\beta = 3$ to 5 both procedures produced approximately same results.

Units	Auxiliary variable X_i	Variable of interest Y_i
1	269	181
2	231	135
3	247	155
4	269	160
5	217	107
6	292	178
7	239	133
8	286	179
9	291	157

Source: Simulation

Table 6: Distribution of the auxiliary variable and variable of interest (Simulated Data II)

Units	MP ($\alpha = 5, \beta = -5$)	BSP ($\alpha = 5, \beta = 2$)
1	0.8883272	0.8899424
2	0.8779062	0.8819528
3	0.8806778	0.8847322
4	0.8883272	0.8899424
5	0.8765553	0.8800803
6	0.9052728	0.8975757
7	0.8790849	0.8832489
8	0.8995969	0.8953362
9	0.9042518	0.8971893
Total	8	8

Source: Researchers' Computation

Table 7: Probability of inclusion, π_i (Data II)

Units Sampled	MP ($\alpha = 5, \beta = -5$)	BSP ($\alpha = 5, \beta = 2$)
1, 2	0.7662334	0.7718952
1, 3	0.7690049	0.7746745
1, 4	0.7766543	0.7798848
1, 5	0.7648824	0.7700227
1, 6	0.7935999	0.7875181
1, 7	0.7674121	0.7731913
1, 8	0.7879241	0.7852785
1, 9	0.792579	0.7871317
2, 3	0.758584	0.766685
2, 4	0.7662334	0.7718952
2, 5	0.7544614	0.7620331
2, 6	0.783179	0.7795285
2, 7	0.7569911	0.7652017
2, 8	0.7775031	0.7772889
2, 9	0.782158	0.7791421
3, 4	0.7690049	0.7746745
3, 5	0.757233	0.7648124
3, 6	0.7859505	0.7823078
3, 7	0.7597627	0.7679811
3, 8	0.7802747	0.7800683
3, 9	0.7849296	0.7819214
4, 5	0.7648824	0.7700227
4, 6	0.7935999	0.7875181
4, 7	0.7674121	0.7731913

Comparison of Some Sample Selection Procedures In Probability Proportional to Size

4, 8	0.7879241	0.7852785
4, 9	0.792579	0.7871317
5, 6	0.781828	0.777656
5, 7	0.7556402	0.7633292
5, 8	0.7761522	0.7754164
5, 9	0.7808071	0.7772695
6, 7	0.7843577	0.7808246
6, 8	0.8048697	0.7929118
6, 9	0.8095246	0.7947649
7, 8	0.7786819	0.7785851
7, 9	0.7833368	0.7804382
8, 9	0.8038488	0.7925254
Total	28	28

Source: Researchers' Computation

Table 8: Joint Probability of inclusion, π_{ij} (Data II)

Units Sampled	MP($\alpha = 5, \beta = -5$)	BSP ($\alpha = 5, \beta = 2$)
1, 2	44.44794	42.60931
1, 3	13.34622	13.01336
1, 4	8.973464	8.648178
1, 5	120.257	114.6985
1, 6	0.67725	0.3682295
1, 7	48.42361	46.33615
1, 8	0.3245606	0.1755114
1, 9	12.24676	11.58868
2, 3	9.486971	8.687751
2, 4	12.34451	12.0146
2, 5	20.08252	18.42073
2, 6	27.11605	31.74834
2, 7	0.1200707	0.1115799
2, 8	32.21613	34.89723
2, 9	5.888667	7.485472
3, 4	0.2931301	0.3454118
3, 5	55.91483	51.47591
3, 6	6.117719	8.065347
3, 7	11.59226	10.61639
3, 8	8.106149	9.458954
3, 9	0.08221214	0.0006265052
4, 5	60.72343	58.07099

4, 6	3.634295	4.912261
4, 7	14.61475	14.17597
4, 8	5.063908	5.948744
4, 9	0.5681649	0.3306472
5, 6	83.14077	92.99529
5, 7	16.87117	15.42603
5, 8	94.45606	99.35184
5, 9	40.23543	45.25024
6, 7	30.00914	34.89165
6, 8	0.06537867	0.03517318
6, 9	5.927783	7.206062
7, 8	35.45017	38.21455
7, 9	7.369872	9.164628
8, 9	7.687571	8.441207
Total	843.8759	865.1815

Source: Researchers' Computation

Table 9: Least variance estimates (Data II)

		α									
		-5		-4		-3		-2		-1	
β		MP	BSP	MP	BSP	MP	BSP	MP	BS P	MP	BS P
	-5	902	926	894	924	886	920	879	915	872	909
	-4	909	926	902	924	894	920	886	915	879	909
	-3	915	926	909	924	902	920	894	915	886	909
	-2	920	926	915	924	908	920	901	915	894	909
	-1	923	926	919	924	913	920	907	915	900	909
	1	926	925	923	923	920	919	915	914	908	908
	2	926	926	924	924	920	920	915	915	909	909
	3	926	926	924	924	920	920	915	915	909	909
	4	926	926	924	924	920	920	915	915	909	909
	5	926	926	924	924	920	920	915	915	909	909

Source: Researchers' Computation

Table 10: Comparison of variances using PP II and BSP (Simulated Data II)

Comparison of Some Sample Selection Procedures In Probability Proportional to Size

		α									
		1		2		3		4		5	
β		MP	BSP	MP	BSP	MP	BSP	MP	BSP	MP	BSP
	-5	860	894	854	886	850	879	847	872	844	865
	-4	865	894	860	886	854	879	850	872	847	865
	-3	872	894	865	886	859	879	854	872	850	865
	-2	879	894	872	887	865	879	859	872	854	865
	-1	884	895	877	887	870	879	864	872	858	866
	1	894	893	886	885	879	893	871	885	865	878
	2	894	894	886	886	879	879	872	872	865	865
	3	894	894	886	886	879	879	872	872	865	865
	4	894	894	886	886	879	879	872	872	865	865
	5	894	894	886	886	879	879	872	872	865	865

Source: Researchers' Computation

Table 10 continues

	Least variances Estimates	
Data Sets	MP	BSP
Data I	479	523
Data II	844	865

Source: Researchers' Computation

Table 11: The least variance estimates using MP and BSP

As the value of α remains fixed and β increases for the MP, the variance estimates remain approximately the same as seen in Tables 5 and 10. On the other hand, as the value of β remains fixed and α increases, the variance estimates decreases for the MP. This shows that here, β could be fixed while α varies. Table 11 shows the least variance estimates obtained using MP and BSP for the two simulated data sets.

In conclusion, we recommend the use of the proposed sample selection procedure in unequal probability sampling without replacement.

References

1. Alodat, N. A. On unequal probability sampling without replacement sample Size 2. *Int. J. Open Problems Comp. Math.*, **2(1)**: 109 – 112 (2009).
2. Basit, A. and Shahbaz, M. Q. A general class of selection procedures a modified Murthy estimators. *Journal of Statistics*, **14**: 109 – 112 (2007).

3. Horvitz, D.G. and Thompson, D. J. A generalization of sampling without replacement from a finite population. *Jour. Amer. Statist. Ass.*, **47**: 663 – 685 (1952).
4. Sen, A.R. On the estimate of the variance in sampling with varying Probability. *Jour. Ind. Sos. Agric. Statist.* **5**: 119 – 129 (1953).
5. Shahbaz, M. Q. Sampling with unequal probabilities without replacement. *Unpublished Ph. D thesis* (2003).
6. Shahbaz, M. Q., Shahbaz, M. and Hanif, M. On new selection procedures for unequal probabililty sampling. *Int.J. Open Problems Compt. Math.*, **4(1)**: 177 – 182 (2011)..
7. Yates, F. and Grundy, P.M. Selection without replacement from within strata with probability proportional to size. *J. Roy. Stat. Soc. B*, **15**: 153 – 161 (1953).