International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 6, Issue 1, January 2017 **ISSN 2319 - 4847**

Sample selection with probability proportional to size sampling using SAS and R software

NobinChandra Paul

Ph.D. Scholar, Indian Agricultural Statistics Research Institute, New Delhi, India

ABSTRACT

In simple random sampling every unit in the population has the equal chance of being included in the sample but if the sampling units vary in their size then simple random sampling does not take into account the importance of larger units in the population contradiction to this probability proportional to size sampling scheme gives more probability of inclusion to larger units as compared to the smaller units in the sample. Therefore, Probability proportional to size sampling gives more efficient estimator as compared to equal probability sampling scheme. In this article, my aim is to select a sample of units on the basis of its size by using SAS and R software.

Keywords: Probability Proportional to Size (PPS), SAS,PROC SURVEYSELECT

1.INTRODUCTION

In simple random sampling (SRS) probability of selection of every units in the population is equal but when sampling units are varying in their size if in that case SRS is being used for sample selection then unexpected result may arises as some of the more important (larger) units may not be included in the sample. Under this circumstance, a more precise estimator is used which takes into account the size of every sampling units and probability is assigned based on their sizes which is known as probability proportional to size (PPS) sampling [1]. This paper mainly focused on sample selection procedure based on the size of sampling units by using SAS and R software.
Let us consider Y is the variable under study which is the holdings in a village and let X is an auxiliary variable which indicates number of fields in a particular holding. Here, holdings are selected with proportion to their number of fields. So, here number of fields can be considered as measure of size while selecting holdings from a village [1].

2. PPS SAMPLING WITH REPLACEMENT

In PPS sampling with replacement, the probability of selection of a unit will not change and the probability of drawing a specified unit is same at any stage [10]. Generally two method is used for sample selection in PPS sampling with or without replacement which includes cumulative total method and lahiri's method but here my aim is to select sample by using statistical software (SAS and R). Therefore, i mainly focused on the programs which is being used to draw sample PPS with or without replacement.

2.1 PPS Sampling with Replacement using SAS

SURVEYSELECT procedure is used for sample selection. PROC SURVEYSELECT provides methods for both equal probability and PPS sampling. To select a sample with PROC SURVEYSELECT, we need to input a SAS data set that contains list of units from which the sample is to be selected. We need to also specify the sample selection method, desired sample size and other selection parameters [5]. This procedure produces an output data set that contains the selected units, their selection probabilities and their sampling weights. General syntax of PROC SURVEYSELECT is given below

PROC SURVEYSELECT options;
STRATA variables</options>;
CONTROL variables;
SIZE variable;
ID variables;

The PROC SURVEYSELECT statement starts the procedure and identifies input and output data sets. It also specifies the selection method, the sample size.

The SIZE statement identifies the variable that contains the size measure of the sampling units.The remaining statements are optional. The STRATA statement identifies a variable or set of variables that stratify the input data set. The CONTROL statement identifies variables for ordering units within strata. The ID statement identifies variables to copy from the input data set to the output data set of selected units [5].

2.1.1 SAS code for sample selection PPS with Replacement

I used SAS version 9.4 (Copyright © 2002-2013) of Sas Institute Inc., Cary, NC, USA for sample selection [8]. SAS code is given on the basis of a practical (hypothetical) example of a village which has 10 holdings consisting of a specified number of fields which is given in Table 1.

Table1: The data set of the fields in a holding of a particular village [1].

Holdings (Y)	Fields (X_i)
1	40
2	45
3	35
4	50
5	60
6	25
7	30
8	28
9	38
10	44

SAS code for PPS with replacement is given below,

```
data ppswr;
input Holdings Fields;
Cards;
1 40
2 45
3 35
4 50
5 60
6 25
7 30
8 28
9 38
10 44
;
run;
proc surveyselect data=ppswr method=pps_wr sampsize=4 out=sample_wr;
size Fields;
Run;
Proc Print;
run;
```

Figure1 SAS code for PPS sampling with replacement.

In PROC SURVEYSELECT statement, **data=** specify the input data set, **method=** specify the sample selection method which is **pps_wr** for PPS sampling with replacement, **out=** specify output data sets and it also specifies the sample size by using **sampsize**option [5]. Here, the size measure is Fields.

The output of the following program is given below,

The SAS System

The SURVEYSELECT Procedure

Selection Method	PPS, With Replacement
Size Measure	Fields

Input Data Set	PPSWR
Random Number Seed	506103001
Sample Size	4
Output Data Set	SAMPLE_WR

The SAS System

Obs	Holdings	Fields	NumberHits	ExpectedHits	SamplingWeight
1	1	40	1	0.40506	2.46875
2	2	45	1	0.45570	2.19444
3	7	30	1	0.30380	3.29167
4	10	44	1	0.44557	2.24432

Figure 2 Output of the program for PPS sampling with replacement in SAS

Therefore, from the output we can see that holding number 1, 2, 7 and 10 are selected with corresponding sampling weight 2.46875, 2.19444, 3.29167 and 2.24432.

2.2 PPS Sampling with Replacement using R Software

I used R software version 3.1.3 (Copyright © 2015) of the R foundation for statistical computing for sample selection [9]. For sample selection, PPS package developed by Jack G. Gambino is used. The package consists of several functions for selecting a sample from a finite population in such a way that the probability of a unit being selected is proportional to its size, hence the name is PPS [2], [6].

So, the R code for PPS with replacement based on the same data set which is in Table 1 is given below,

```
> Install.packages('pps')
> Library(pps)
> data <-read.csv( file.choose(),header = T)
> sizes <-c(40,45,35,50,60,25,30,28,38,44)
> ppswr(sizes,4)
```

First pps package is installed. After installing pps package, data is loaded using **read.csv(file.choose(),header=T)** functions. Data set is saved in comma separated value (CSV) format and **header=T** means data set contain header. The function **ppswr(sizes,4)** selects a sample of 4 units with replacement, with the probability of selection of each unit proportional to its size, where sizes is a vector of the sizes of the population units and 4 is the sample size. In our example, sizes are the number of fields in a particular holding.

So, the output of the code is given below,

```
> install.packages('pps')
--- Please select a CRAN mirror for use in this session ---
trying URL 'http://ftp.iitm.ac.in/cran/bin/windows/contrib/3.1/pps_0.94.zip'
Content type 'application/zip' length 133875 bytes (130 KB)
opened URL
downloaded 130 KB

package 'pps' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\NOBIN CH PAUL\AppData\Local\Temp\RtmpGE3ad4\downloaded_packages
> library(pps)
> data<-read.csv(file.choose(),header=T)
> sizes<-c(40,45,35,50,60,25,30,28,38,44)
> ppswr(sizes,4)
[1] 10  4  2 10
>
```

Figure 3 Output for PPS sampling with replacement using the above R code.

Therefore, holding number 10 is selected two times and holding number 4 and 2 selected once.

3. PPS SAMPLING WITHOUT REPLACEMENT

In PPS sampling without replacement initial probabilities of selection are unequal and the probability of drawing a specified unit of the population at a given draw changes with the draw. Sample selection under PPS without replacement sampling using statistical software are given in following sections.

3.1PPS Sampling without Replacement using SAS

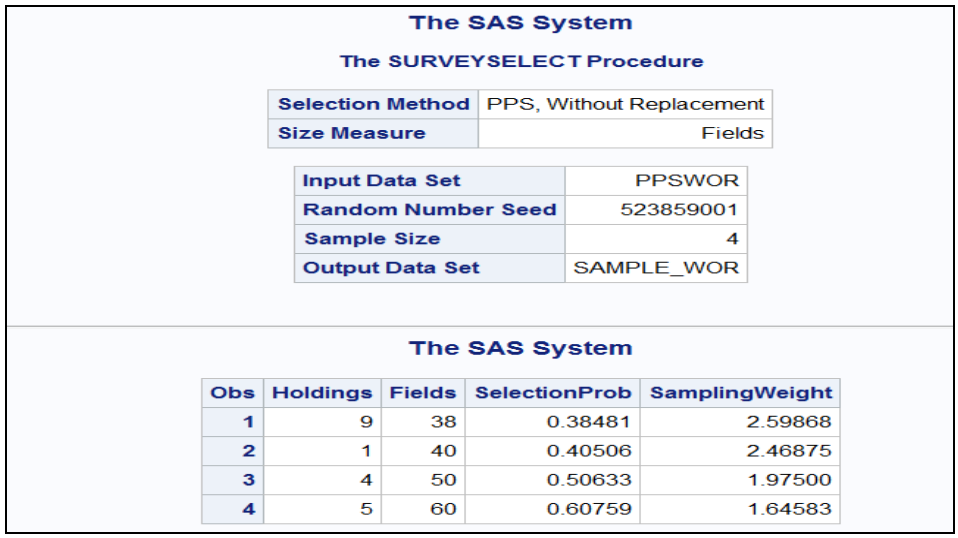
The same PROC SURVEYSELECT procedure is used in PPS sampling without replacement but the only difference in the **method=** options, here we need to specify **method=pps,** PROC SURVEYSELECT selects units with PPS and without replacement [5].

SAS code is given below for the same data set of Table 1.

```
data ppswor;
input Holdings Fields;
Cards;
1 40
2 45
3 35
4 50
5 60
6 25
7 30
8 28
9 38
10 44
;
run;
proc surveyselect data=ppswor method=pps sampsize=4 out=sample_wor;
size Fields;
Run;
Proc Print;
run;
```

Figure 4 SAS code for PPS sampling without replacement.

The output of the following program is given below,



The SAS System

The SURVEYSELECT Procedure

Selection Method	PPS, Without Replacement
Size Measure	Fields

Input Data Set	PPSWOR
Random Number Seed	523859001
Sample Size	4
Output Data Set	SAMPLE_WOR

The SAS System

Obs	Holdings	Fields	SelectionProb	SamplingWeight
1	9	38	0.38481	2.59868
2	1	40	0.40506	2.46875
3	4	50	0.50633	1.97500
4	5	60	0.60759	1.64583

Figure 5 Output of the above code based on PPS sampling without replacement in SAS

Therefore, the output indicates holding number 9, 1, 4 and 5 is selected with corresponding sampling weight 2.59868, 2.46875, 1.97500 and 1.64583.

3.2 PPS Sampling without Replacement using R Software

In R, the function**sampford** is used to select a PPS sample without replacement using sampford's method. The function **sampfordpi**computes the corresponding inclusion probability π_i and joint inclusion probability π_{ij}. This **sampfordpi** function creates a matrix with the π_i along the diagonals and π_{ij} along the off diagonals [2], [3] and [4].

So, the R code of PPS without replacement using the same data set of Table 1 is given below,

```
> Install.packages('pps')
> Library(pps)
> data <-read.csv(file.choose(),header = T)
> sizes <-c(40,45,35,50,60,25,30,28,38,44)
> sampford(sizes,4)
> sampfordpi(sizes,4)
```

Output of the code is given below,

```
> install.packages('pps')
--- Please select a CRAN mirror for use in this session ---
trying URL 'http://ftp.iitm.ac.in/cran/bin/windows/contrib/3.1/pps_0.94.zip'
Content type 'application/zip' length 133875 bytes (130 KB)
opened URL
downloaded 130 KB

package 'pps' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\NOBIN CH PAUL\AppData\Local\Temp\Rtmpg1y8M4\downloaded_packages
> library(pps)
> data<-read.csv(file.choose(),header=T)
> sizes<-c(40,45,35,50,60,25,30,28,38,44)
> sampford(sizes,4)
[1] 5 9 1 6
```

Figure 6 Output of sample selection of PPS without replacement using R software

```
> sampfordpi(sizes,4)
             [,1]       [,2]       [,3]      [,4]      [,5]       [,6]
 [1,] 0.40506329 0.15515245 0.11671636 0.1754186 0.2180532 0.08085364
 [2,] 0.15515245 0.45569620 0.13372161 0.2001716 0.2480759 0.09284644
 [3,] 0.11671636 0.13372161 0.35443038 0.1513903 0.1887226 0.06936558
 [4,] 0.17541856 0.20017157 0.15139028 0.5063291 0.2787920 0.10537006
 [5,] 0.21805320 0.24807592 0.18872258 0.2787920 0.6075949 0.13204234
 [6,] 0.08085364 0.09284644 0.06936558 0.1053701 0.1320423 0.25316456
 [7,] 0.09848569 0.11296829 0.08458081 0.1280552 0.1600636 0.05834719
 [8,] 0.09136379 0.10484653 0.07843100 0.1189059 0.1487808 0.05406269
 [9,] 0.12796183 0.14649620 0.11009488 0.1657217 0.2062387 0.07619967
[10,] 0.15118436 0.17280959 0.13026804 0.1951621 0.2420157 0.09040607
             [,7]       [,8]       [,9]      [,10]
 [1,] 0.09848569 0.09136379 0.12796183 0.15118436
 [2,] 0.11296829 0.10484653 0.14649620 0.17280959
 [3,] 0.08458081 0.07843100 0.11009488 0.13026804
 [4,] 0.12805524 0.11890585 0.16572168 0.19516210
 [5,] 0.16006364 0.14878082 0.20623866 0.24201565
 [6,] 0.05834719 0.05406269 0.07619967 0.09040607
 [7,] 0.30379747 0.06601113 0.09285636 0.11002405
 [8,] 0.06601113 0.28354430 0.08612659 0.10210450
 [9,] 0.09285636 0.08612659 0.38481013 0.14273450
[10,] 0.11002405 0.10210450 0.14273450 0.44556962
```

Figure 7 Output of matrix with the inclusion probabilities along the diagonal and joint probabilities along the off diagonals.

So, the output indicates holding number 5,9,1 and 6 is selected based on the field sizes.

4. CONCLUSIONS

This paper provides programs for sample selection for both PPS with and without replacement sampling based on a hypothetical example given in Table 1 by using statistical software (SAS and R). It is known to us that Hansen and Hurwitz first introduced probability proportional to size (PPS) sampling [7], [11]. There are different procedures have been developed by statistician for sample selection with and without replacement and they have found that PPS sampling scheme is more efficient for sample selection as well as parameter estimation when sampling units varies in their sizes [7]. Finally, it is clear from the result that the above programs can efficiently be utilized for sample selection under probability proportional to size with and without replacement sampling scheme.

REFERENCES

[1]. Daroga Singh and F. S. Chaudhary, Theory and Analysis of Sample Survey Designs, New Age International (P) Ltd., Publishers, New Delhi,1986.

[2]. J. Gambino, "Users's Guide to R functions for PPS Sampling", 2003. [Online].Availvable:http://hbanaszak.mjr.uw.edu.pl/Sampling/Software/R/Gambino_2003_User%92s%20guide%20to%20R%20functions%20for%20PPS%20sampling.pdf. [Accessed: Jan. 04, 2017].

[3]. W. G. Cochran, Sampling Techniques, John Wiley and Sons, New York, 1977.

[4]. Sampford, M. R., "On Sampling without replacement with unequal probabilities of selection,"Biometrika, vol. 54, pp.499-513, 1967.

[5]. SAS Institute Inc., **"**SAS/STAT 9.2 User's Guide The SURVEYSELECT Procedure (Book Expart)", Cary, NC,USA,2008.[Online].Available:https://support.sas.com/documentation/cdl/en/statugsurveyselect/61839/PDF/default/statugsurveyselect.pdf. [Accessed: Jan. 04, 2017].

[6]. J. Gambino, "Functions for PPS Sampling", 2015. [Online]. Available: https://cran.r-project.org/web/packages/pps/pps.pdf. [Accessed: Jan. 04, 2017].

[7]. MaskurulAlam, SharminAkter Sumy and Yasin Ali Parh, "Selection of the Samples with Probability Proportional to Size," Science Journal of Applied Mathematics and Statistics, vol. 3, no. 5, pp. 230-233, 2015.

[8]. SAS version 9.4 (Copyright © 2002-2013) of Sas Institute Inc., Cary, NC, USA.

[9]. R version3.1.3, Copyright © 2015. The R foundation for Statistical computing.

[10]. Dr. Shalabh, "Varying probability sampling", [Online]. Available: http://home.iitk.ac.in/~shalab/sampling/chapter7-sampling-varying-probability-sampling.pdf. [Accessed: Jan. 04, 2017].

[11]. Hansen, M. H. and Hurwitz, W. N., "On the theory of sampling from a finite population," Annals of Mathematical Statistics, vol. 14, pp. 333-362, 1943.