International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-7, Issue-6, June 2021
ISSN: 2395-3470
www.ijseas.com

Systematic Sampling Scheme for Sample Surveys Using Electronic Spreadsheet

Edike Nnamdi
Department of Mathematics and Statistics, Ambrose Alli University, Ekpoma.
nnamdiedike@gmail.com

Umar Shehu Salisu
Department of Statistics, The federal Polytechnic, Kduna, Kaduna, State, Nigeria
umarss83@yahoo.com

Braimah Joseph Odunayo
Department of Mathematics and Statistics, Ambrose Alli University, Ekpoma.
ojbraimah2014@gmail.com

Agbedeyi Odimientimi Desmond
Department of mathematics and statistics, Delta state polytechnic, ogwashi-uku
desmondagbedeyi@gmail.com

Abstract

Systematic sampling is perhaps one of the most adopted sampling techniques in sample surveys such as household surveys, price collection surveys and even health surveys. A major procedure involved using this technique is the computation yielding the *serial*number of the units to be included in the sample. This is quite time consuming especially when a fairly large sample is required.This paper, therefore presents the design and implementation of an electronic spreadsheet which enhances sample selection using the systematic sampling technique. The spreadsheet was designed using Microsoft Excel and an illustration on how it works was explicitly provided. The principles of the systematic sampling technique were implemented using Microsoft Excel formulas and syntaxes and could possibly be included into the application add-in tools.

Keywords: spreadsheet, sample survey, Microsoft Excel, systematic sampling, formula

1. Introduction

Systematic sampling is a statistical method involving the selection of units from an ordered sampling frame. The procedure begins by selecting an element from the list at random and then every k^{th} element in the frame is selected where *k* is the sampling interval.

This sampling scheme has a long tradition in survey sampling (e.g., Madow and Madow 1944, Madow 1949, 1953 in Li-Chun, 2008). It is also known as the "every *kth*rule" especially when applied to a list of units.When the ordering of the list of units is conceivably uncorrelated with the variable of interest, or contains at most a mild stratification effect, the systematic sampling is generally considered as a convenient substitute for simple random sampling "with little expectation of a gain in precision"(Cochran 1977). In situations where auxiliary information is available for partial ordering of the population, it is more natural to compare systematic sampling with stratified random sampling.

Given that we wish to obtain a sample of size *n* from a population of size *N*, we obtain *k* by the relation $k = \frac{N}{n}$, then the first element is sampled randomly either by using the raffle method, table of random numbers, etc. this is known as the random start (RS)(Ken, B., 2004). We then proceed by selecting every k^{th} element in the frame until we get to the end of the frame. The random start say, *a* must be less or equal to the sampling interval,*k*. The procedure described above yields a sample of size *n* where each element in the sampling frame has a known and equal probability of being included in the sample. The systematic sampling is usually suitable where the given population is logically homogenous since the sample units are uniformly distributed over the population.

In demographic surveys, sampling is usually done in phases, the first phase being listing of population units e.g. heads of households (for households surveys), this serves as a sampling frame which helps the investigator to work out a good sample selection procedure. The sample selection phase is very important because any error in this phase will eventually lead to biasness of survey outcomes and wrong inference about the population under study. Most of these errors in sampling could be traced to the enumerator due to the tediousness of the procedures and computations involved in sample selection, especially when the population is considerably large. This paper is therefore aimed at implementing this procedure using a computer spreadsheet application (with particular reference to Microsoft Excel), which will significantly reduce the tediousness; saves time and improve accuracy of the process.

The systematic sampling technique is particularly more convenient than the simple random sampling. It also ensures equal probability of each unit being included in the sample. In this method of sampling, the first unit is selected with the help of random numbers and the remaining units are selected automatically according to a predetermined pattern (Shalabh, 2021)

Suppose the units in the population are numbered 1 to*N* in some order. Suppose further that *N*is expressible as a product of two integers n and, so that *N = nk*

According to Shalabh, (20…), To draw a sample of size, *n:*

i. Select a random number between 1 and*k*
ii. Suppose it is*i*
iii. Select the first unit whose serial number is *i*
iv. Select every*k*th unit after *i*th unit.
v. Sample will contain *i, i+k, i+2k, …,i+(n-1)k*serial number units.

So first unit is selected at random and other units are selected systematically. This systematic sample is called*k*th systematic sample and*k* termed as sampling interval. This is also known as linear systematic sampling.

This sampling scheme has the advantage of being easier to draw a sample and often easier to execute without mistakes. This is also more advantageous when the drawing is done in fields and offices as there

may be substantial saving in time. The cost is low and the selection of units is simple. Much less training is needed for surveyors to collect units through systematic sampling. Furthermore, The systematic sample is spread more evenly over the population. So no large part will fail to be represented in the sample. The sample is evenly spread over the population.

A good number of sampling techniques can be carried out in Microsoft Excel using built-in functions, for example, the simple random sampling (SRS) technique can be achieved using built-in functions, this is done using the analysis toolpak under data analysis group of the Data menu. One of the functions similar to the systematic random sampling is the periodic sampling. However, the periodic sampling in Microsoft Excel is different in a number of ways from the systematic sampling technique. These are:

1. The periodic sampling scheme in MsExcel gives no room for a random start.
2. The period (sampling interval) is automatically the starting point. This is not the case in systematic sampling.
3. The number of sample and population cannot be specified. This makes Excel unable to compute realistic sampling interval.
4. The period (sampling interval) must be an integer. This is not always the case is systematic sampling.

However, the systematic sampling scheme can be achieved in entirety using some user-defined functions and facilities provided in the spreadsheet. This is what this paper is aimed to achieve.

2. Methods

This paper utilized the facilities and functions available in spreadsheet packages (with particular reference to Microsoft Excel) to implement the systematic random sampling procedure. This would be very useful in carrying out sampling during sample surveys such as household surveys, demographic and health surveys etc. it would be very helpful to field statisticians who are directly involved in data collection during field surveys. It will eliminate the stress of manual computation and reduce error inherent in sample selection as a result of computation involved in sampling procedures. The systematic sampling scheme is particularly considered in this paper due to its wide range of application by field statisticians in carrying out various field surveys.

2.1 Algorithm for Systematic Sampling

1. Assign numbers 1 to N to the population elements using the three-digit format, i.e. the first element is assigned 001, the second, 002 and so on.
2. Calculate the sampling interval$=\frac{N}{n}$, where N is the population size and n the desired sample size. There may arise cases where the sampling interval k is fractional, in such cases the decimal part of the selected sample number is truncated.
3. Using raffle draw technique or table of random numbers, select the first unit with number a, where $a \leq k$. If the selected unit has number$a > k$, discard and repeat the process.
4. Add the sampling interval k to the first in order to obtain the second number, i.e. $a + k$. Repeat this step to get $a + 2k, a + 3k, a + 4k, \ldots, a + ck$, where $a + ck \leq N$

5. The units/elements corresponding to the numbers obtained in step (4) above are included in the sample.

2.2 The Spreadsheet

The spreadsheet for the implementation of the above procedure consists of two(2) worksheets; the first is the *feeder* worksheet, it contains a table of three columns which include the serial number (SN), the unit number and the unit identity (this could be the name of head of household, for household survey).The second is the *sample* worksheet, it shows the selected units, which has a table of four columns. This could be printed for administrative convenience. The feeder worksheet allows for data entry. Both worksheets consist of additional cells provided for entry of other vital information such as the population size, desired sample size and the random start. The sampling interval is automatically computed by the spreadsheet as soon as the required variables are entered. The spreadsheet is equally designed to prompt the user with an error message in cases where he erroneously entered invalid data, for instance, where the random start, *a*isgreater than the sampling interval, *k* or where the sample size *n* is greater than the population size *N*.

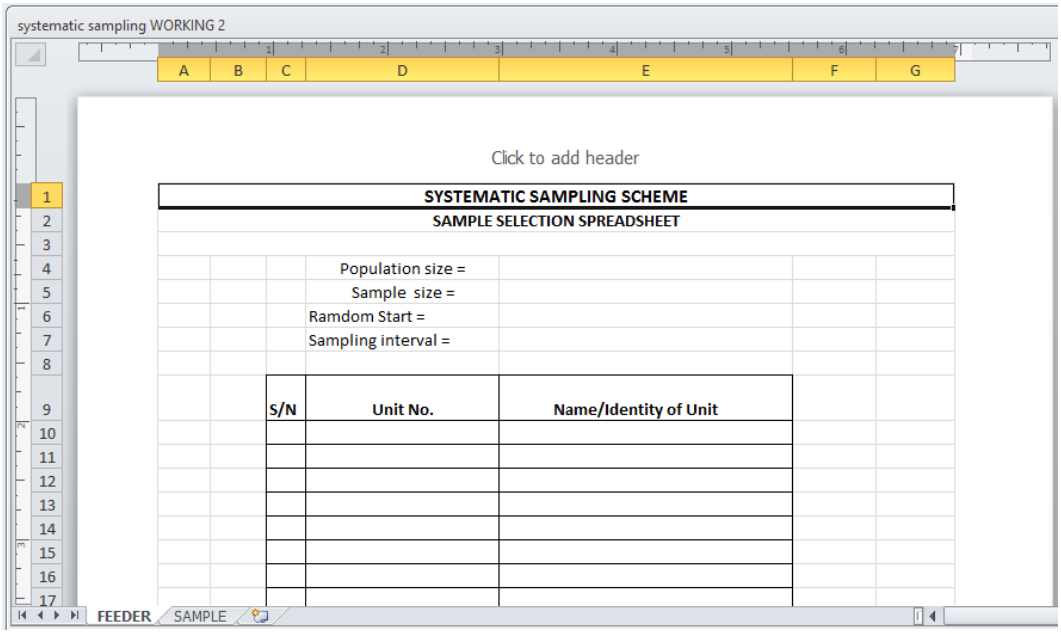


Figure 1: The Feeder Worksheet

In the figure above, the size of the population from which the sample is to be drawn is entered in cell E4, while the sample size, entered in cell E5. Hence cell E7 will contain the formula "**=E4/E5**". This gives the sampling interval. Furthermore, for easy identification in the formula, the cells provided for the population size, sample size, random start and sampling interval were named "N", "S", "RS", and "SI" respectively. These names will be used to identify these cells while working on the spreadsheet.

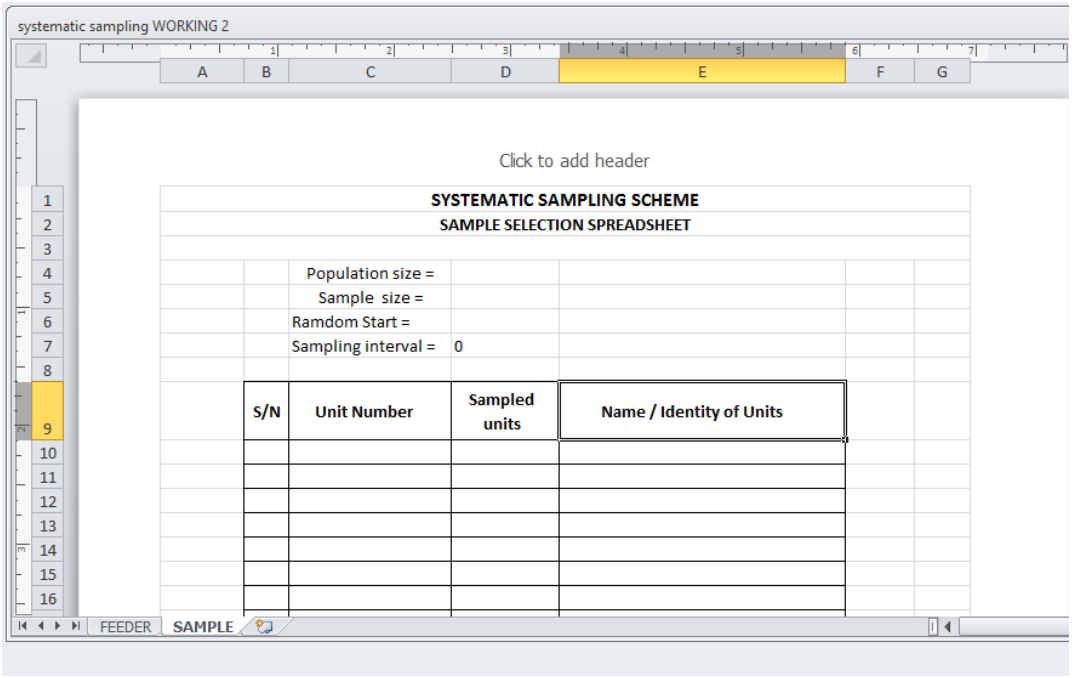


Figure 2: The Sample Worksheet

This worksheet displays the final sample. The selected units are shown in this worksheet alongside their respective identities. To protect the sheet from being altered, the SAMPLE worksheet is protected and does not allow for data entry, every single item displayed here are either imported from the FEEDER worksheet or generated automatically with the use of coded formulas.

2.3 Procedure:

1. On the S/N column (cell B10), we enter the formula "**=IF(FEEDER!C10<=$D$5,FEEDER!C10,"")**". This imports the serial number in the FEEDER worksheet for serial numbers less than or equal to the specified sample size, otherwise the column is left blank.
2. On the first row under Sample values column, we enter the formula "**=IF(FEEDER!RS="","",FEEDER!RS)**". This imports the value of the random start (RS) from the FEEDER worksheet; it is left blank if the cell containing the random start is blank.
3. On the second row under the Sample values column, we enter the formula "**=IF(B11="","",IF(C10="","",C10+FEEDER!SI))**". This formula instructs the spreadsheet to add the value of the sampling interval in FEEDER worksheet to the value of the random start in cell C10, which in turn gives the unit number of the next unit to be included in the sample. However, the cell is left blank if the adjacent cell in serial number column (cell B11) is blank. This ensures that no unit is selected if serial number is not displayed since the serial number truncates at the number of sample size required.

4. On the sampled units column (cell D10), we enter the formula "**=IF(C10="","",TRUNC(C10))**". This formula brings out the actual unit number of the selected units by truncating the result in the adjacent cell (C10). This ensures that the column does not contain numbers with decimal points.
5. On the identity of units column (cell E10), we enter the formula "**=IF(D12="","",VLOOKUP(D12,FEEDER!$C$9:FEEDER!$E$200,3,0))**". This formula looks up the value of the sampled units in the table in FEEDER worksheet and fetches the corresponding identity of the selectedunits.
6. In each of the steps above the use drag button at the bottom right corner of the cell pointer to extend the formula down the column.

2.4 How to use the worksheet

1. Launch the workbook from the Microsoft Excel application.
2. Open the FEEDER worksheet (usually active by default).
3. Enter the population size, the desired sample size and the random start in the spaces provided.
4. Click on the SAMPLE sheet tab to view the selected units.
5. Print out or publish the content of the worksheet to a desirable output media.

3. Illustration

We illustrate the use of the developed worksheet by selecting a sample of size 20 from a population of 120 households. We arbitrarily assigned the unit numbers to each of the population units. The names / initials of heads of households were equally entered in the column provided. All inputs were entered on the feeder worksheet.

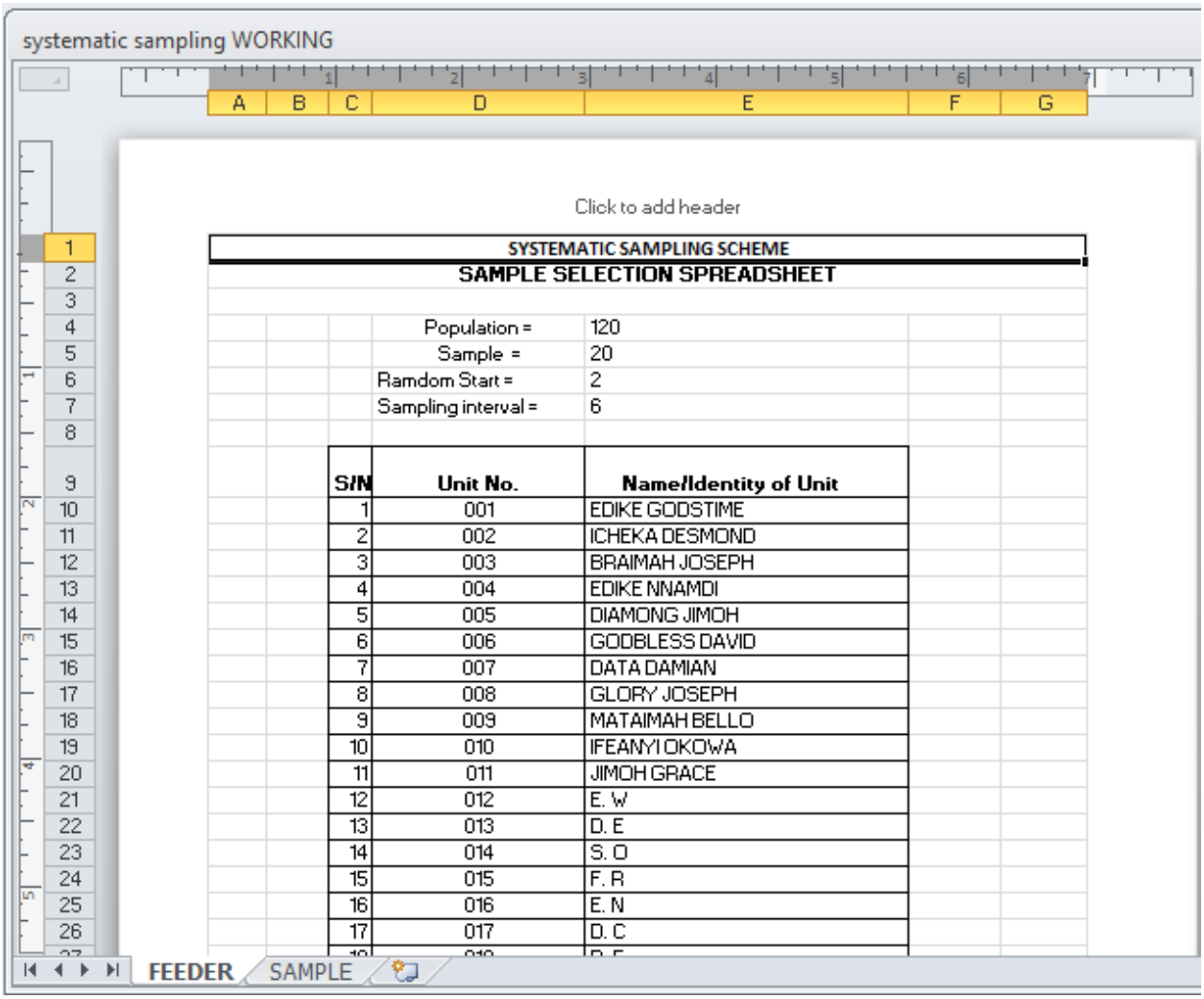
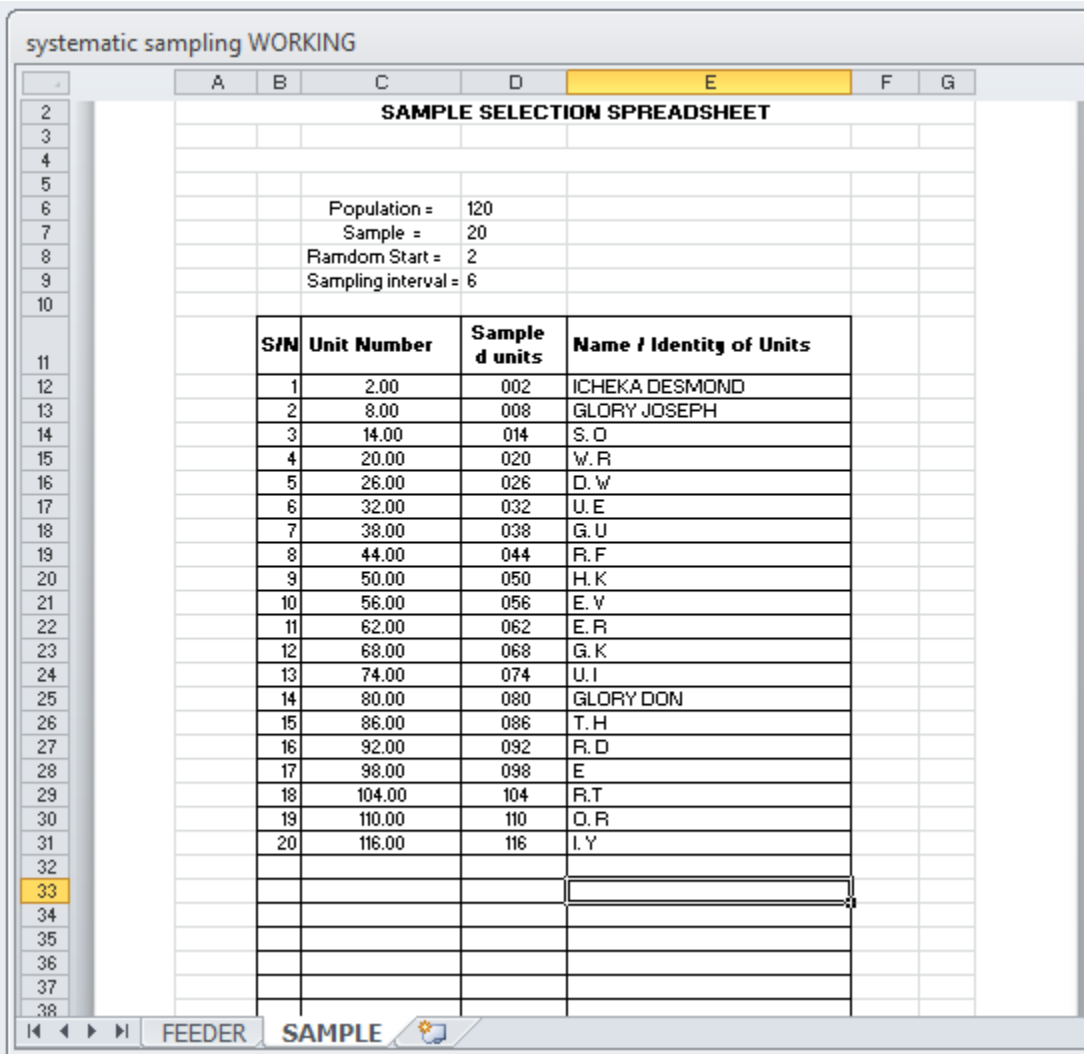


Figure 3: The FEEDER worksheet showing information on the listed population units.

The random start was chosen to be 2 which is of course less than the sampling interval of 6, computed by the spreadsheet. Having entered the required information, we click on the SAMPLE worksheet to see the selected households. The worksheet below shows the selected sample.



systematic sampling WORKING

SAMPLE SELECTION SPREADSHEET

Population =	120
Sample =	20
Ramdom Start =	2
Sampling interval =	6

S/N	Unit Number	Sampled units	Name / Identity of Units
1	2.00	002	ICHEKA DESMOND
2	8.00	008	GLORY JOSEPH
3	14.00	014	S. O
4	20.00	020	W. R
5	26.00	026	D. W
6	32.00	032	U. E
7	38.00	038	G. U
8	44.00	044	R. F
9	50.00	050	H. K
10	56.00	056	E. V
11	62.00	062	E. R
12	68.00	068	G. K
13	74.00	074	U. I
14	80.00	080	GLORY DON
15	86.00	086	T. H
16	92.00	092	R. D
17	98.00	098	E
18	104.00	104	R.T
19	110.00	110	O. R
20	116.00	116	I. Y

FEEDER SAMPLE

Figure 4: The SAMPLE worksheet showing the selected units from the population

3.1 Error notification

Two major errors are likely to occur while using the spreadsheet. The first case is when the user erroneously enters a sample size greater than the population size and the second is when he enters a random start greater than the sampling interval. These are conventional errors that violate the principles of sampling and of course, systematic sampling scheme in particular.

When such error occur, the spreadsheet flags off an error message on a red strip indicating the type of error committed in each case. Figures 5 and 6 below show the error messages flagged by the spreadsheet in each case.

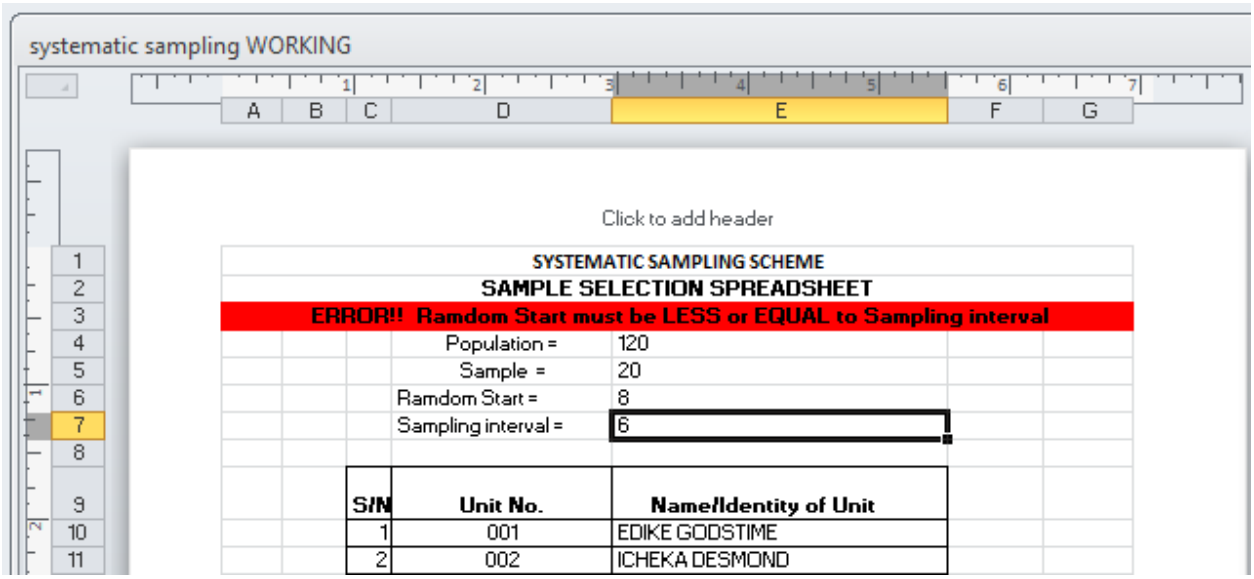


Figure 5: Showing error notification when Random Start is greater than Sampling Interval

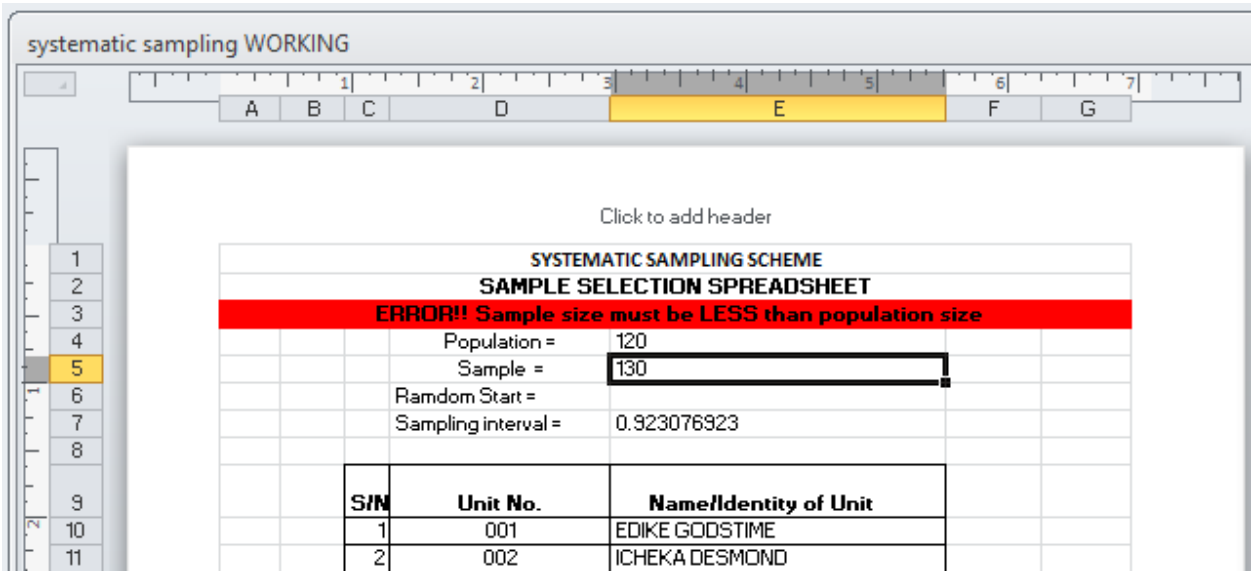


Figure 6: Showing error notification when sample size is greater than population size

4. Conclusion

This paper presents theimplementation of the systematic scheme using an electronic spreadsheet. It demonstrates the usefulness of a spreadsheet in carrying out sample selection using the systematic sampling scheme. The principles of systematic sampling scheme were briefly reviewed and implemented using the Microsoft Excel spreadsheet. The spreadsheet formula and syntaxes used were clearly shown and an illustration of its usage was also demonstrated. This paper will no doubt provide solution to the problem of

sample selection often encountered by field officers and enumerators during sample surveys, especially those involving the use of the systematic sampling scheme.

REFERENCES

Li-Chun, Z. (2008).On Some Common Practices of Systematic Sampling.Journal of Official Statistics, Vol. 24, No. 4, 2008, pp. 557–569

Ken, Black.(2004). Business Statistics for Comtemporary Decision Making.Fourth (Wiley Student Edition for India) ed. Wiley-India.ISBN 978-81-265-0809-9.

Cochran, W.G. (1977). Sampling Techniques, (3rd ed.). New York: John Wiley and Sons.

Shalabh (2021). Lecture Note on systematic sampling. Lecture note 11 on Shalabh Lecture Series, available on https://home.iitk.ac.in/~shalab/spsampling.htm. Retrieved on 3rd May 2021.