Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.4, No.6, 2014

On the Selection of Samples in Probability Proportional to Size Sampling: Cumulative Relative Frequency Method

Faruq Abdulla
Student, Department of Statistics, Islamic University, Kushtia-7003, Bangladesh,
Email: faruqiustat09mnil@gmail.com

Md. Moyazem Hossain (Corresponding author)
Department of Statistics, Islamic University, Kushtia-7003, Bangladesh,
E-mail: mmhrs.iustat@gmail.com

Md. Mahabubur Rahman
Department of Statistics, Islamic University, Kushtia-7003, Bangladesh,
E-mail: sagorju151@gmail.com

Abstract

Generally in the sense that, the unit with large size contain more ancillary information than the unit with smaller size. So when samples from different sized subgroups or units are used and sampling is taken with the same probability, the chances of selecting a member from a large group are less than selecting a member from a smaller group although here the chances of selecting a member from a large group will be greater than selecting a member from a smaller group. That is it is clear that, the probability of selecting a unit is positively proportional to its size. The aim of this paper is to propose a method of selecting samples in probability proportional to size. This method uses relative frequency to select samples in probability proportional to size. Comparatively it takes less time and easy to apply than Cumulative Total Method and Lahiri's Method.

Keywords: Probability Proportional to Size (PPS) Sampling, Cumulative Total Method, Lahiri's Method, Cumulative Relative Frequency Method.

1. Introduction

Sampling is a research method where subgroups or units are selected from a larger group known as a target population. The subgroups or samples are studied. If the sample is correctly chosen then the results can be used to represent the target population. In almost all practical situations the size of the sampling units whose are included in the sampling frame are vary from one to another. Generally in the sense that, the unit with large size contain more ancillary information than the unit with smaller size. So when samples from different sized subgroups or units are used and sampling is taken with the same probability, the chances of selecting a member from a large group are less than selecting a member from a smaller group although here the chances of selecting a member from a large group will be greater than selecting a member from a smaller group. That is it is clear that, the probability of selecting a unit is positively proportional to its size. This is known as Probability Proportional to Size (PPS) sampling. If the selected unit with associated size in the sample is reconsidered in the sampling frame then it is known as PPS sampling with replacement but if the selected unit with associated size in the sample is ignored from the sampling frame then it is known as PPS sampling without sampling. There exist numerous methods for selecting a fixed size sample with probability proportional to size (PPS), with replacement. Traditionally, Simple Random Sampling (SRS) has been regarded as the method-of-choice because all elements in the population have an equal chance to get picked and hence, reducing selection bias and

selection through SRS is relatively convenient when sampling frame is readily available.

There are a number of ways to randomly select units. If the units are arranged in a list which is assigned identification codes, say the integers between 1 and N we can use either of the following two methods:

Lottery: In a lottery, the ID codes of all possible units in the target population are written on a ballot (e.g. a separate piece of paper or a separate ping-pong ball) and placed into a bowl (sometimes referred to as an ``urn''). The ballots are shuffled (randomized), and then ballots are drawn from the bowl. Drawing can occur in two ways. One is without replacing a drawn ballot before the second ballot is drawn (selection without replacement). Or with replacement and reshuffling of the contents of the bowl before the second ballot is drawn (selection with replacement). Note that in selection with replacement, an individual can be selected twice in the same sample. But this procedure of numbering of units on marble/disc/piece of paper and selecting one after reshuffling becomes cumbersome when the population and desired sample size is large and also human bias and prejudice may also creep in this method.

Random Number Table: Numerous random tables are available. Tables of random numbers are tables of the digits $0, 1, 2, \ldots, 9$, each digit having an equal chance of selection at any draw. Among the larger tables are those published by the Rand Corporation (1955) – 1 million digits – and by Kendall and Smith (1938) – 100,000 digits. A table of random numbers can be used to randomly select numbers between 1 and N. The procedure is as follows:

- *Select random page in random number table*

Open to the first page of the random number table. Place a pencil in one hand, close your eyes and use the pencil to point to a digit on the table. Use that digit to determine which page of the random number table to turn to. So, for example, if there are 9 pages in the random number table and your pencil lands on a 7, you would turn to page 7 for the next step. If the selected number is not a valid page number, read digits down the column until you come to the first valid page number and then proceed using that page number. If the number of pages in the random number table is between 10 and 99, you would read two digits at a time from the table, again reading down the column until you reach a valid page number. If you come to the end of the column without obtaining a valid page number, proceed to the upper left top of the page and continue reading. In the very unlikely event you don't find a valid page number on the first page, proceed to the second page.

- *Select random starting point on selected page*

Open to the randomly selected page. Again, with a pencil in one hand, close your eyes and use the pencil to point to a digit on the table.

- *Create list of unit ID numbers for sample*

If N is between 1 and 10, you will read one digit, using 0 to indicate 10. If N is between 00 and 99, you will read two digits, using 00 to indicate 100, etc. Any number which is not a valid sample unit ID number is simply discarded and you simply proceed to read the next set of digits from the next row on the table. Continue reading set of digits and recording acceptable numbers until one has selected the number of units needed. Thus if you need $n = 25$ units in the sample, one would continue reading sets of digits from the table until 25 acceptable ID numbers were listed.

The aim of this paper is to propose a method of selecting sample in probability proportional to size. This method uses relative frequency to select sample in probability proportional to size. Comparatively it takes less time and easy to apply than Cumulative Total Method and Lahiri's Method. The organization of this paper is as follows: Section 2 discussed the literature review. The existing and proposed methods of selecting random sample in probability proportional to size were in section 3. Concluding remarks are discussed in section 4.

2. Literature Review

Since 1943 when Hansen and Hurwitz first introduced the use of probability proportional to size (PPS) sampling, a number of procedures for selecting samples without replacement have been developed. Many of them are reviewed and compared in Brewer and Hanif (1983). 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). Selection with PPS means that, on the average, a primary sampling unit (psu), which is, for example, 5 times as large as another will be in the sample 5 times as frequently as the other psu. It might appear, at first, that this would introduce a bias in the sample result with some psu's overrepresented and others under-represented - and it would, in fact, if no special attention were given to the varying probabilities in the estimating or subsampling procedures (Hansen, Hurwitz, and Madow, 1953a, p342). One of the advantages gained by the use of PPS sampling is that: for many common populations, and with a fixed number of primary units in the sample, a smaller variance will be obtained by sampling with probability proportionate to size than by sampling with equal probability. (Hansen, et al. l953b).

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 pi 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 nonlinear 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.

Tillé (1996) presents a new PPS, without replacement sampling procedure. The general idea of this procedure is to begin with the universe and eliminate one unit at a time until the number of remaining units is equal to the desired sample size. The set of remaining units then becomes the original sample. Although Tillé does not discuss applying his procedure to the sample expansion problem, the method of doing so is essentially immediate provided a record had been kept of the order of elimination of the units not in the original sample. A key result in this paper is that it is also always possible to perform a sample expansion with the required properties when the original sample had been selected using Tillé's method and no record had been kept of the order of elimination of the units.

3. Methods of Selecting Samples in Probability Proportional Size

3.1 Cumulative Total Method

Let the size of the i^{th} unit be $X_i\ (i=1,2,\cdots,N)$, the total being $X=\sum_{i}^{N} X_i$. We associate the numbers 1 to X_1 with the first unit, the numbers (X_1+1) to (X_1+X_2) with the second unit, and so on. A number k is chosen at random from 1 to X and the unit with which this number is associated is selected. Clearly, the i^{th} unit in the population is being selected with a probability proportional to X_i. If a sample of size n is required, the procedure is repeated n times with replacement of the units selected. This procedure of selection is known as the cumulative total method for the method needs cumulation of the unit sizes.

3.2 Lahiri's Method

Lahiri's method consists in selecting a number at random between 1 to N and noting down the unit with the corresponding serial number, provisionally. Another random number is then chosen between 1 to M, where M is the maximum size of the N units of the population. If the second random number is smaller than the size of the unit provisionally selected, the unit is selected into the sample. If not, the entire procedure is repeated until a unit is finally selected. For selecting a sample of n units, the procedure is to be repeated until n units are selected.

3.3 Cumulative Relative Frequency Method

Let the size of the i th unit be $X_i\ (i=1,2,\cdots,N)$, the cumulative total of the i th unit being $CT_i=\sum_{r=1}^{i} X_r\ (i=1,2,\cdots,N)$, and the total being $X=\sum_{i=1}^{N} X_i$. We obtain the corresponding cumulative relative frequency of the i th unit as $CRF_i=\frac{CT_i}{X}(i=1,2,\cdots,N)$. Now we construct intervals as 0 to CRF_1 for the first unit, CRF_1 to CRF_2 for the second unit, CRF_2 to CRF_3 for the third unit, and so on. A number k is chosen at random from uniform distribution over the range 0 to 1 or from a random number table from 0 to 1 and indentify the interval where k lies. Then the corresponding unit of the selected interval is selected. If the chosen number falls on the boundary point, it is suggested to consider exclusive method of classification. If a sample of size n is required, the procedure is repeated n times with replacement of the units. This procedure of selecting sample is known as the *cumulative relative frequency* method.

Example: Suppose a village has 10 holdings consisting of 50, 30, 45, 25, 40, 26, 24, 35, 28 and 27 fields, respectively. Now we select a sample of 4 holdings with the replacement method and with probability proportional to the number of fields in the holding. In order to do this, first we calculate the following table:

Sl. No. of Holdings	Size (X_i)	Cumulative Size	Cumulative Relative Frequency	Interval
1	20	20	0.0625	0.0000 - 0.0625
2	30	50	0.1562	0.0625 - 0.1562
3	45	95	0.2969	0.1562 - 0.2969
4	25	120	0.3750	0.2969 - 0.3750
5	40	160	0.5000	0.3750 - 0.5000
6	26	186	0.5812	0.5000 - 0.5812
7	44	230	0.7188	0.5812 - 0.7188
8	35	265	0.8281	0.7188 - 0.8281
9	28	293	0.9156	0.8281 - 0.9156
10	27	320	1.0000	0.9156 - 1.0000

To select a holding, a random number is drawn from 0 to 1 with the help of uniform distribution over the range 0 to 1 or from a random number table. Suppose the selected random number selected is 0.4678. It can be seen from the cumulative relative frequency that the number is lies in the interval 0.3750 - 0.500, and the 5^{th} holding is selected since it the corresponding holding of 0.4678 is 5. Similarly, we select three more random numbers. Suppose these numbers are 0.832, 0.2743 and 0.654. Then the holdings selected corresponding to these random numbers are 9^{th}, 3^{rd} and 7^{th} respectively. Hence, a sample of 4 holdings selected with the replacement method and with probability proportional to size will contain the 3^{rd}, 5^{th}, 7^{th} and 9^{th}.

5. Conclusion

Since 1943 when Hansen and Hurwitz first introduced the use of probability proportional to size (PPS) sampling, a number of procedures for selecting samples without replacement have been developed. 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. This paper proposes a method of selecting samples in probability proportional to size which use relative frequency. It is shown that this method takes less time and easy to apply. Thus it recommended to apply the relative frequency method for selecting samples in probability proportional to size.

REFERENCES

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.

Brewer, K. R. W. and Hanif, M. (1983). *Sampling with unequal probabilities*, New York: Springer-Verlag.

Chaudhuri A (2010). *Essentials of Survey Sampling*, PHI Learning Private Limited, New Delhi.

Ekaette, IE, (2008). *A class of alternative estimators for Multicharacteristics 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.

Hansen, Morris H., William N. Hurwitz and William G. Madow., (1953a). *Sample survey Methods and Theory, Volume 1- Methods and Applications*, New York. John Wiley and Sons.

Hansen, Morris H., William N. Hurwitz and William G. Madow., (l953b). *Sample Survey Methods and Theory, Volume 2- Theory*, New York. John Wiley and Sons.

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

Kendall, M. G., and Smith, B. B., (1938). Randomness and random sampling numbers, *Jour. Roy. Stat. Soc.*, 101, 147-166.

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

Rand Corporation (1955). *A Million Random Digits*, Free Press, Glencoe, III.

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

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

Tillé, Y. (1996). An Elimination Procedure for Unequal Probability Sampling Without Replacement, *Biometrika*, 83, 238-241.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Recent conferences: http://www.iiste.org/conference/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar