THINK INDIA JOURNAL

ISSN:0971-1260
Vol-22- Issue-12-November-2019

Analysis of Cluster and Multistage Sampling: A Survey on Higher Education Discontinuity in Rural Areas using Cluster Sampling

Dr. Vijay Kumar Garg
Associate Professor, CSE
Lovely Professional University
Jalandhar, Punjab
vijay.garg@lpu.co.in

Abstract-This paper has tried to emphasize on a vital aspect of research that is sampling. Sampling, if done in a proper way then one can decide the overall success or failure of any research project. This paper is an honest attempt in trying to shed light on two of the very efficient methods of sampling (cluster and multistage) which can lead to provide us the most accurate results if done correctly. The paper clearly tries to show the importance of the cluster sampling technique by conducting an in-depth literature review and then by illustrating the exact impact it can have on a research by conducting a small survey by using the Cluster sampling technique only, where an effort has been made to find out reasons for discontinuity in higher education in rural areas along with their causes. The results are shown with the help of tables and figures which describe that due to various reasons the ratio of female respondents is greater than male respondents in the discontinuity in higher education after matriculation.

Keywords- Cluster Sampling, Discontinuity, Higher Education, Matriculation Students

1. INTRODUCTION

Sampling is used of finding units (e.g., individuals and organization) from a inhabitants to enable us to find out the results back to the inhabitants from which they were taken by analyzing the sample. Every calculation tests one or more of an observable entity's properties (weight, position, etc.) specified to differentiate independent entities or individuals. A sample is part of a target population that is deliberately chosen to represent the population. The sampling frame is the set of elements that actually draw the sample from. Sampling is, in truth, nothing but the right population list [17].

1.1 Sampling Process: In sampling process, various activities are performed as shown in Fig. 1.
1.1.1 Test Model Types: Sampling is essentially divided into two types:

Probability Sampling: The choice of each unit in population has equal chances of being chosen as a sampleunit in aprobability sampling and this probability is calculated perfectly. It comprises many sampling techniques as shown in Fig. 2.

ISSN:0971-1260
Vol-22- Issue-12-November-2019

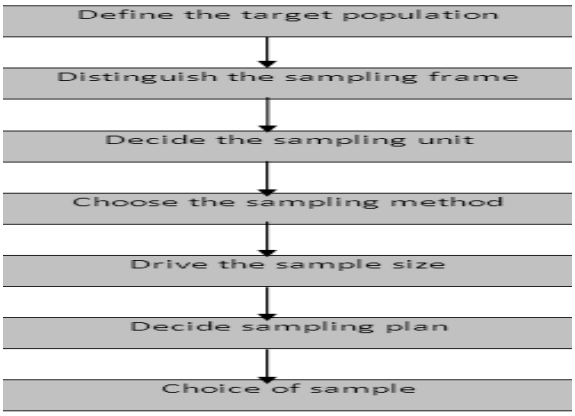

Fig.1 Steps for Sampling Process

Example:

Suppose we want 7,500 students from across the country to be selected. In such a case first we select District, say that 30 out of 600 districts are selected from across the country and then:

I Stage – Cities: Suppose five of the 30 districts are selected.

II Stage – Schools: from each city, 10 schools are chosen.

III Stage – Students: 50 students from each school were picked.

We may use stratified sampling in stage I.

We may use cluster sampling in stage II.

We may have simple random sampling in stage III.

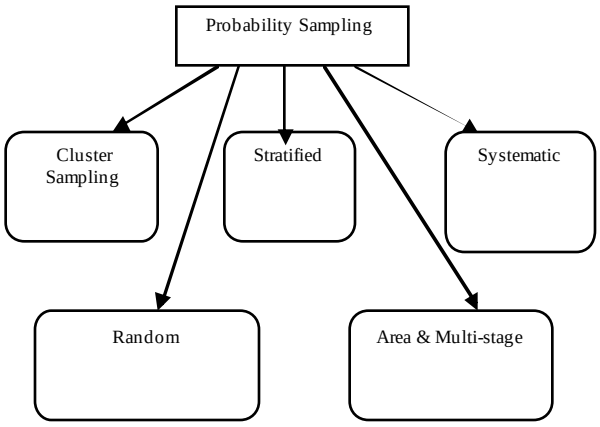

Fig.2 Classification of Probability Sampling

Non-probabilitySampling: This technique have the chances of selecting units are unequal or negligible in a non-probability sampling, with almost no chance of being selected as a sample unit. The choice of

ISSN:0971-1260
Vol-22- Issue-12-November-2019

the elements for the inclusion in the sample is dependent on researcher. The various sampling techniques come under non-probability sampling are shown in Fig. 3.

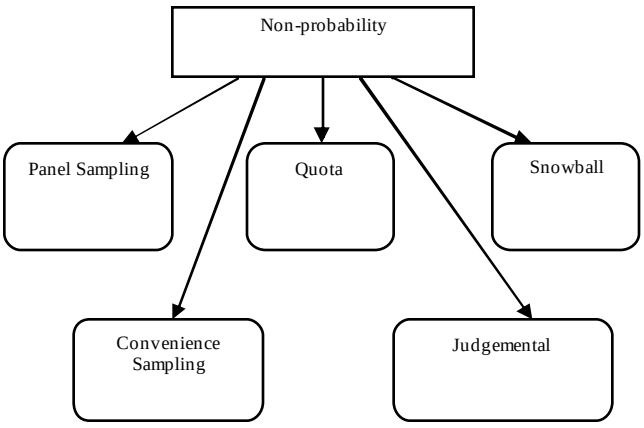

Fig.3 Classification of Non-probability Sampling

This paper is based on in-depth analysis of cluster and multi-stage sampling. So, first have a brief look on these sampling techniques.

1.2 History of Sampling: The sampling method first used on test survey of unemployment in 1937 Census Bureau. Detailed questions about the population can be asked without any overhead costs due to the sampling method. Sampling and estimation techniques are also used by the Census Bureau to measure net coverage in the decennial census. After that net coverage is compared with the census results nationwide. In the preparation, creation and operation of the processes of its numerical schemes, sampling techniques were used by the census Bureau to ensure their quality along with its effectiveness [14].

1.3 Cluster sampling: Consider the sampling of the cluster in the selection of samples. Suppose we have a population of 40,000 units with 500 units to choose from. Selecting a sample of this size using the Random Number table is a very complicated process. Now we can easily select two sample clusters (2X250=500) by using cluster sampling, dividing the entire population into 160 clusters of 250 units each. From the above situation it is clearly stated that the following steps will be taken in group sampling.

• The community will be divided into groups.
• Select a simple random sample of just a few clusters.
• All units are being studied in the selected cluster.

Advantages:

- Cost-effective in group / cluster selection of respondents.
- Reduces administrative cost and travel expenses.
- Does not include a sampling frame listing all the target population components.

ISSN:0971-1260
Vol-22- Issue-12-November-2019

1.4 Area Sampling: This is a special form of sampling of the cluster i.e. a kind of a probability based sampling. Sample area is taken with the help of map. The procedure that follows in area sampling is to partition the large areas into small areas. A random selection is made within each of the area selected, after that a subsample of locality is taken and then monitored.

1.5 Multi-stage Sampling: Sampling is done in several stages as the name implies. For cluster / stratified layout, this is used. The following is a brief view of the two-stage sampling. A newly opened shopping mall's management is calling for new membership. The entire collective was sent during the first round so that those interested could enter. The second round, after registration observes how many are interested in registering for various activities provided by shopping mall management, such as food corner, entertainment, theaters etc. You can stratify the interested joined after collecting this data.

Example: To find out which mental health is more important education and income. A survey was conducted at Santiago, Chile using three stage cluster sampling. In which all participants were taken for private household as random sample of persons aged 16-65.As a result, less employment (values 2.44, 95 percent having intervals between 1.50 and 3.97), a recent decline in salary (value 2.14, 1.70 to 2.70), and poor living place (odds ratio 1.53, 1.05 to 2.23) were the only socio economic status factors that remain significant associate with increased prevalence of common mental disorder after small change. The prevalence of common mental disorder among people with unqualified manual employment, overpopulated housing, and lower per capita income was also higher, but these correlations removed after correction for other interpretive and confusing variables. Understanding the impact on mental health of socio-economic factors requires research in both poor and rich countries [15].

TABLE-I

DESCRIPTIVE ANALYSIS OF CLUSTER SAMPLING

Type Used	Name of Work	Technique Used	Brief Description
Imbalanced Data Set (Over Sampling Method)	Borderline-SMOTE [1]	Synthetic Minority Over-sampling Technique(SMOTE)	Imbalancing means distribution of classes in data sets are not balanced like as we can not accurately predict the how many intruders are available at one time to attack on network, failure of servers in a particular organization. Imbalancing can be in between classes where some classes can have more probability to contain more examples than other classes. Te other is an imbalance within the class, where some subgroups have less description than other. Class having more examples known as majority class and which have less examples called as minority class. This research paper provides solution to imbalance problem in data mining that was previously addressed by SMOTE. This experiment shows two new minority Over Sampling (i.e. sampling is being done at the highest level to adjust the class distribution) approaches using borderline-SMOTE1 and borderline-SMOTE2 that have better F-value and True Positive rate than previous SMOTE and random over-sampling methods.
Imbalanced Cluster Set (Under	Cluster based under-sampling [2]	Down-sampling based on clustering and	The term imbalancing is already known to all that distribution of classes is not balanced and in most of the cases classifiers assume that incoming information belongs to majority group rather than minority class.

ISSN:0971-1260
Vol-22- Issue-12-November-2019

Sampling Method)		distances between samples	This work was based on the solution of imbalanced class distribution problem by using cluster-based down-sampling approach and to boost the classification accuracy for minority class (i.e. too few sample size is being taken).The author presents two NearMiss-2, Random selection and down-sampling methods for comparison with the previous studies and it is found that SBC(Sampling Based Cluster) approach has better prediction results and prediction power in comparison to others.
Dynamic test Cluster Sampling	A Dynamic Test Cluster Sampling [3]	Execution-spectra-based sampling (ESBS)	Cluster filtering is used to save human effort as a sample selection strategy by increasin g test size and identifying total failures. In the cluster filtering process, cluster sampling techniques play an important role. A sampling technique called execution-spectra-based sampling is used in this paper which differs from current sampling strategies because execution-spectra-based sampling takes test subjects from each cluster iteratively. ESBS selects the test case in each iteration cycle that has the highest probability of being a failed test. For each court, the expectation is calculated on the basis of knowledge from past passed execution spectra and failed test cases from the same cluster described. The experimental result shows that in most situations, ESBS is more successful in detec ting errors than current sampling techniques.
Adaptive Cluster Sampling (ACS)	Improvement of Low Level Bark Beetle Damage Estimates with Adaptive Cluster Sampling[4]	Relative efficiency estimator	Adaptive Cluster Sampling (ACS) implies that enlarging the plot once a target element is found on the initial plot and that target objects can be rare tree or shrub species etc. in case of forest inventory. ACS is adaptable to specific situations that means, final design that is implemented is not completely predictable but depends also on what is being found out there.This conditional adaptation of the design makes estimation difficult, because the selection probability is then obviously a conditional probability and the selection probability of a specific element depends also on the proximity of other elements.
Clusters Partial Sampling	Developing an appropriateness Model for Supportive Supervision in the Educational System of Iran[5]	Factorial Analysis and One Sample T Test	The purpose of this paper is to identify the components of supportive supervision, building an appropriate model for supportive supervision and also determine the degree of appropriateness of proposed model. For this a questionnaire was prepared and cluster sampling method and within clusters partial sampling was used and to answer the questions the literature was reviewed and global studies were surveyed and some components were extracted and identified separately. To test the components, the collected data were disaggregated through statistical methods with high frequency percentage and through Factorial Analysis and One Sample T Test and accuracy of the questionnaire was calculated as %93 and the total mean of 8/9 out of 10.

TABLE-2

DESCRIPTIVE ANALYSIS OF MULTI-STAGE SAMPLING

Type Used	Name of Work	Technique Used	Brief Description
Sampling Structure Conversion	Multistage Sampling Structure Conversion of Video Signals[6]	Lattice Theory Descending and Ascending Lattice Chains Multi-Dimensional Sampling Structure Conversion	This paper incorporates multi-stage transformation of the sample structure to the multi-dimensional case. Because of the relationships between transfer bandwidth and filter order for 1–D FIR filters, the benefit of this structure conversion improves efficiency. Both are dependent on each other, which means that if the order of the FIR filter is higher, the transition bandwidth is lower. In multi-dimensional cases, bands of transition are multi-dimensional (possibly irregular) regions and not simple intervals, as in 1-D cases. The main purpose is to convert the video signals structure that can improve system and visual rendition characteristics. These concepts were applied to conversion of video format.

ISSN:0971-1260
Vol-22- Issue-12-November-2019

Two-Stage Case-Control Study	Estimating Equations and Multistage Sampling Designs [7]	HorvitzThompson Estimator Mean Method	Multi-stage sampling has demonstrated cost efficiency in this paper by minimizing the variance of a parameter estimate or maximizing the units with some desirable attribute, subject to a fixed total budget constraint.For any combination of Bernoulli and fixed fraction sampling, these variances apply.
Four-Stage Sampling	Multistage Sampling for Disease Family Registries[8]	Horvitz–Thompson approach Statistical Theory	Multi-stage sample technique can be used to establish an accurate registry of diseases. Samples will be taken on the basis of past taken data at each stage, and for the observation purpose, a subsample would be selected. This technique can be enhanced with respect to the cumulative size of sample and the usage of model parameter estimates to reduce the variance. The Thompson method is appliedon a conventional genetic two-stage test by using estimate group-based parameters for the University of Southern California. The goal of the Breast and Colorectal Cancer Research Cooperative Family Registry includes establishing a family information database for use in both genes.
Multi-Stage Sampling	Patterns for the double use of s moking and che wing tobacco a mong US males findings from r ecent surveys [9]	Complex sample design.	This paper is linked to a study of U.S. males ' dual use of cigarettes a nd smokeless tobacco. To carry out this survey stratified analysis technique is used and also analysis is taken as references from the previous four survey paper. The result was determined by reporting p- values for certain direct correlations of means or proportions, and estimated means for the recorded standard errors and confidence interval was measuredby proportions. The Men who have the high usage of moist stuff they had less choice of smoking as compared to others who take less moist stuff.

TABLE-3
LITERATURE REVIEW ON HIGHER EDUCATION DISCONTINUITY ON DIFFERENT PERSPECTIVE

Name of Work	Technique Used	Brief Description
Student loans: Liquidity constraint and higher education in South Africa[10]	Regression discontinuity design	This paper provides direct demonstration of comparison of different university enrollment rates of South African because students admission depends on university enrollment rate. Due to which sometimes they borrow loan to cover their registration fees. Regression discontinuity model is used for this analysis to assess the fact that loans are issued on the basis of a credit score threshold and also to measure the causal effect of receiving loans. But the graduation rate in higher education falls by more than 20 percentage points in a student loan borrower population due to various credit constraints.
Running in Place: Less Income v/s Dynamicity in Higher Education[11]	Multinomial logistic regression, Simple descriptive statistics	This Paper introduces the problems of low-income students that have high academic achievements and are perfectly matched to enroll in the selective colleges/institutions for admissions and the opposite side wealthy students with not so much strong academic records become successful to enroll for higher education. By using multinomial logistic regression, the researcher proved that That Social Stratification in Institutional Destination has an impact on low-income students to

ISSN:0971-1260
Vol-22- Issue-12-November-2019

		enroll in selective colleges, but organizational and policy initiatives can reverse these trends.
Would military draft hinder higher education enrollment? Proof from countries of the OECD [12]	F-test, descriptive statistics	This paper analyzes the higher education demand effect of compulsory military service. Based on theoretical model and descriptive statistics, it is proved that compulsory military service including duration of service reduces a negative effect in enrollment in higher education.
School tracking access to higher education [13]	Regression discontinuity design, Descriptive statistics.	This paper takes advantage of Romania's educational reform to examine the impact of postponing tracking on the percentage of disadvantaged university graduates using a discontinuity regression (RD) design. We show that while students from poor rural areas and parents with less education are significantly more likely to complete an academic course and become eligible for university applications.

II. SURVEY BASED ON CLUSTER SAMPLING

2.1 Goals of the Study: The purpose of this survey is to investigate the reason for discontinuity in Higher Education after matriculation in rural areas. In this analysis, the below mentioned questions were asked:

(a) What is the student readiness rate for the higher test?
(b) What are the various stumbling blocks that they are facing to continue the higher education?
(c) What is the level and encouragement of both parents and teachers to study higher?
(d) Have they gained sufficient information to pursue the higher study?

2.2 Participants and Settings: The bigger challenge in this survey is to find out the size of the cluster set. To calculate sample size of students, need to complete the following seven steps [16]:
Use the following formula to calculate the sampling interval:

Interval of Sampling = total cumulative population ÷ number of required clusters
= 1400 ÷ 3 = 466

TABLE-4
CUMULATED TOTAL POPULATION

Village	Nomination size	Cumulative
1	400	400
2	250	650
3	300	950
4	200	1150

ISSN:0971-1260
Vol-22- Issue-12-November-2019

5	250	1400

To estimate the nomination in each school Nomination size = Total number of schools× average number of students in a school

Select a random number that is equal to or less than the interval of sampling. We have to choo se a random number from 1 to 466. Let's say that the number picked is 260.Look at the table now. Choose that cluster whose cumulative value exceeds the random number 260. In our sample, in Village 1, where the total value is 400, the first cluster would be found.

Add the random number of the sampling interval. According to this resultant would be 466 + 260 = 726.Choose the village that only exceeds this number. The second cluster will therefore be located in the third village. Identify the location of each subsequent cluster by adding the sampling interval to the previous cluster number. Stop when as many clusters as you need to be found. Now, 3rd cluster will be computed as 726 + 466 = 1192 and next exceeded value is 1400.So according to calculation village 5th will be considered as third cluster. According to the previous steps, total 3 clusters are made and randomly we can choose any of the clusters that will meet our requirements.

Calculate the sample size using the formula:

$$\text{Sample size} = \frac{4 \times \text{ratio} \times (1\text{-ratio}) \times \text{effect of design}}{\text{Failure margin} \times \text{Failure margin}}$$

2.3 Proportion: Make a rough estimate of the proportion of students experiencing more challenges when pursuing higher education using any available information. This is called the anticipated proportion. According to Fig. 4.that proportion is 90% i.e., 0.9.

2.4 Effect of Design: When using cluster sampling, respondents are not picked in the same neighborhood completely independently of the other respondents. Because measurements of sample size are usually based on simple random samples, to account for a large design effect, the sample size must be increased. Using experience from other cluster surveys, we usually allow most variables to have a design effect of 2.0, what we need to double the sample size compared to a simple random sample.

2.5 Margin of error: That determines how close your estimate should be to the actual rate. A fair margin of error is plus or minus 5 percentage points for national aim tracking.

Sample Size = (4 x 0.9 x 0.1 x 2) / (0.05 x 0.05)

= 288

ISSN:0971-1260
Vol-22- Issue-12-November-2019

Two hundred and fifty–six students (132 male, 124 female) from five schools in rural area of north zone of Punjab participated in the survey that were approximately equal to our sample size and cluster size. All participants were matric students who were expected to be very familiar with their higher study future. This sample was made up of 14 to 17-year-old male and female students who will complete their next year's matriculation exam. The selection of the five schools was guided by the fact that the students were comparable in a fundamental way, although the schools were different depending on location. Students from different schools were taught according to their minor and major subjects by their respective teachers. When the researcher went round to ask if they would like to participate in a study that would inquire about their keen interest in higher study, the participants in this study were invited to their classes. The researchers then told the students that at a time they would be encouraged to respond to a self-administered questionnaire by those who were willing to participate. It was informed to prospective participants that there would be no incentive to take part in the study. Participation in this study was completely voluntary and at any time participants were free to withdraw from the study.

2.6 Tools: The main instrument was a 10-item questionnaire that was self-administered. The questionnaire has been split into two sections. The first section requested each respondent's bio-data and the second section searched for data on various factors that helped persuade the higher study.

2.7 Questionnaire: Different questions for the questionnaire are drawn up in accordance for this study purpose. Questions were drawn up in such a way that the information required could be obtained without providing the respondents with any difficulties. These questions were linked as:

- Interest to persue the higher education.
- Knowledge regarding higher education.
- Teachers/Parents encouragement.
- Affect of fee structure.
- Personal skills enhancement.

2.8 Data Analysis: Bar graphs, tables and percentages were used to analyze data collected using the questionnaire. The advantage of using this method is to make it easy for others to understand the information.

III. RESULTS

ISSN:0971-1260
Vol-22- Issue-12-November-2019

The main objective of this research is to find out why there is discontinuity in rural higher education.Fig.4 shows the numbers of males and females students in schools who were facing problems and those who had no problem in any context. Relatively, more females (n=124, 90 %) than males (n=132, 88%) were facing problems while continuing higher education. Table-5 shows the factors that influencing the discontinuity order in higher education and number/percentage of nominations (male and female) that had participated in the survey. Fig.5 shows the ratio of various factors that affect the students for pursuing the higher education. It is clear that fee structure for higher education approximate equally affect male and female respondents in rural areas. Females have less knowledge about their future and also have restrictions from their parents to go out beyond their region as compared to males.

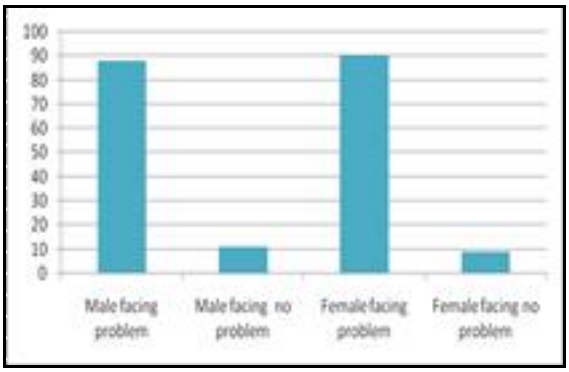

Fig. 4 Ratio of respondents towards discontinuity in higher education

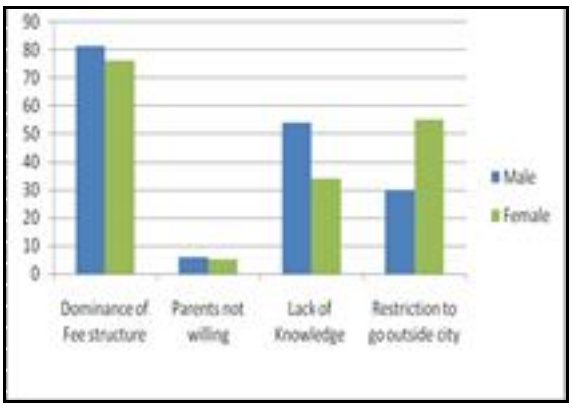

Fig. 5 Factors affecting higher education in context of male and female ratio

TABLE-5
FACTORS AFFECTING HIGHER EDUCATION TO CONTINUE

ISSN:0971-1260
Vol-22- Issue-12-November-2019

Factors	Male		Female	
	No. of Nominations	Percentage of Nominations	No. of Nominations	Percentage of Nominations
Dominace of Fee structure	81	47	76	45
Parents not willing	6	3	5	3
Lack of Knowledge	54	32	34	20
Restriction to go outside city	30	18	55	32
Total	171	100	170	100

IV. CONCLUSION

In this paper, to summarize the current research on discontinuity in higher education, we conducted a study of previous studies published in journals and conferences. The findings of this study offer insights into the complex causes of discontinuity in rural areas of higher education.

Higher education discontinuity is greater in rural areas relative to urban areas. According to our survey some points that are the main contributor towards discontinuity in Higher Education are (i) Lack of knowledge for the higher study, (ii) Willingness of the parents are not involved to indulge for higher study, (iii) Cost of the university/institute is not affordable due to which some of the students has to borrow costly loans, (iv) Sex partiality ratio becomes high in case of female because in rural areas due to the lack of knowledge, parents do not allow their wards to go outside that is the main reason for the female to discontinue the Higher Education.

For overall survey, participants were involved from the various rural schools under the north region of Punjab. To calculate the overall figures that are the major cause for the discontinuity in Higher Education, Cluster size and approximate sample size was calculated. After the analysis it is found that there are only 10% students who have no problem in continuing the Higher Study while the rest of 90% have major concern regarding the disappointment from the institute side and discouragement from the parent's side. From the observation, it is found that female students facing more problems to continue the Higher Education due to various dominance factors. So, to overcome this, at the institute/university level some points should be reserved so that these factors should not become a barrier for students especially in rural areas.

V. REFERENCES

[1] Bing-Huan, Mao Hui Han and Wen-Yuan Wang,"Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning", ICIC , pp. 878 – 887, 2005.

ISSN:0971-1260
Vol-22- Issue-12-November-2019

[2] Yue-Shi Lee,Show-Jane Yen,"Cluster-based sampling approaches for imbalanced data distributions", Expert Systems with Applications, vol.36, pp.5718–5727, 2009.

[3] Zhihong Zhao, Chen Zhang, Yuming Zhou, Shali Yan and Zhenyu Chen, "A Dynamic Test Cluster Sampling Strategy by Leveraging Execution Spectra Information", ICST, pp.147-154, 2010.

[4] Michael A. Wulder ,Sam B. Coggins and Nicholas C. Coops ,"Improvement of Level Bark Beetle Damage Estimates with Adaptive Cluster Sampling", Silva Fennica ,pp.289–301, 2010.

[5] Sohrab Yazdani, Asadollah Khadivi, "Developing Appropriateness Model for Supportive Supervision in the Educational System of Iran", J. Basic. Appl. Sci. Res., vol. 2, pp. 1001-1006, 2012.

[6] G.A. Miana,G.M. Cortelazzo, and R.Manduchi, "Multistage Sampling Structure Conversion of Video Signals", IEEE Transactions on circuits and systems for video technology, vol. 3, 1993.

[7] Alice S. Whittemore, "Multistage Sampling Designs and Estimating Equations", Journal of the Royal Statistical Society", vol. 59, pp. 589-602, 1997.

[8] Kimberly D. Siegmund, Alice S. Whittemore and Duncan C. Thomas, "Multistage Sampling for Disease Family Registries", Journal of the National Cancer Institute, 1999.

[9] Gregory N Connolly,Tomar Hillel, R Alpert and L Scott," Patterns of dual use of cigarettes and smokeless tobacco among US males: findings from national surveys", BMJ Journal Tobacco Control ,pp 104-109, 2010.

[10] Adrien Lorenceau,Marc Gurgand andThomas Melonio ,"Student loans: Liquidity constraint and higher education in South Africa", Paris-Jourdan Sciences Economiques,2011.

[11] Ozan Jaquette,Michael N. Bastedo," Running in Place: Low-Income Students and the Dynamics of Higher Education Stratification", Educational Evaluation and Policy Analysis, vol. 33, pp. 318–339, 2011.

[12] Panu Poutvaara,Katarina Keller and Andreas Wagener, "Does Military Draft Discourage Enrollment in Higher Education? Evidence from OECD Countries", CESIFO, 2009.

[13] Cristian Pop-Eleches,Ofer Malamud, "School tracking and access to higher education among disadvantaged groups", Journal of Public Economics vol. 95, pp.1538–1549, 2011.

[14]Sampling Technique, "http://www.census.gov/history/www/innovatios/data_collection/developing_sampling_techniques.html", Accessed on Oct 25, 2019.

ISSN:0971-1260
Vol-22- Issue-12-November-2019

[15] R Araya, G Lewis, G Rojas, R Fritsch," Education and income: which is more important for mental health," J Epidemiol Community Health; 57:501-505, 2003.

[16] Sampling size, "www.childinfo.org/files/chap04.pdf", Accessed on Oct 16, 2019.

[17] Sampling, "http://en.wikipedia.org/wiki/Sampling", Accessed on Oct 4, 2019.