IJRRS
International Journal of Reliability, Risk
Saftey: Theory and Application

DOI: 10.30699/IJRRS.4.1.8

Vol. 4/ Issue 1/ 2021
pp. 61-67

ORGINAL RESEARCH ARTICLE

Improving Accuracy in Importance Sampling: An Integrated Approach with Fuzzy-Strata Sampling

Mohammad Nadjafi[1*], Adel Najafi ARK[2]

1. Aerospace Research Institute (Ministry of Science, Research and Technology), Tehran, Iran

2. Lecturer and Researcher in Computer Science, Tehran, Iran

Abstract

Several statistical approaches have been developed to analyze the sampling of huge data and information. There are three significant factors for comparison of the strength of these methods that are argued in this paper; the proposed method is a compatible approach to various types of sampling methods and applied to improve the sampling efficiency and decrease uncertainties to reach accuracy in results. In argued methods, each element just belongs to one category and/ or strata, but in our approach, each element includes all groups with one exception that membership values are different. The case study results show that the proposed Fuzzy Strata Sampling (FSS) method better measures uncertainty and accuracy rate than the other existing sampling methods.

Keyword: Sampling method; Stratified Sampling; Fuzzy Approach; Fuzzy-Strata; Accuracy; Uncertainty;

Nomenclature

FSS	Fuzzy Strata Sampling
ME	Margin of Error
SE	Standard Error
SRS	Stratified Random Sampling
μ	Membership value
$\mu_A(x)$	Membership value of x from group A
I	Iteration
N_h	Number of observations within each stratum
n_m	Number of elements in the strata m
N_i	Number of observations in *i*th strata
n_i	Number of elements for *i*th strata
$\bar{n}$	Mean of fuzzy number in each Strata
$\bar{X}_m$	Mean value of *m*th strata

1. Introduction

Improving and handling uncertainty in sampling is a vital component for effective decision making. Uncertainty is insufficiently, explicitly communicated to random sampling methods [1]. The quantification and propagation of uncertainty become essential in precisely those situations where quantitative modeling cannot draw upon extensive historical, statistical or measurement data [2]. Fuzzy logic systems constitute a powerful tool for coping with ubiquitous uncertainty in many engineering applications [3].

The most important point in assessing uncertainty of gained population via sampling is to recognize that all uncertainties are not quantifiable, and therefore they should be separated from the sampling characteristics [4]. In sampling methods, if we are willing to give up some features of random sampling, notably serial independence, then variance reduction techniques may be invoked [5]. A suitable mathematical model for random variables which assume fuzzy values are so-called fuzzy random variables [6].

In standard statistics the combination of observations into an element of the sample space is trivial. But for fuzzy data this is not [7]. Since Zadeh first introduced the concept of a fuzzy set [8], Fuzzy principles have been applied to a huge and diverse range of sciences [9].Survey sampling provides a variety of methods for selecting probability-based random samples [10]. Survey sampling is the process of selecting a probability-based sample from a finite population according to a sample design. You then collect data from these selected units and use them to estimate characteristics of the entire population [11]. The procedure can select a simple random sample or a sample according to a complex multistage sample design that includes stratification, clustering, and unequal probabilities of selection [12]. With probability sampling, each unit in the survey population has a known, positive probability of selection. This property of probability sampling avoids selection bias and enables you to use statistical theory to make valid inferences from the sample to the survey population [13].

The main purpose of most measurements is to enable decisions to be made. The credibility of these decisions

⁎m.nadjafi@ari.ac.ir

Received: 2021.08.02, Accepted: 2021.10. 11, Online ISSN: 2676-3346

depends on knowledge about the uncertainty of the measurement results. Uncertainty in measurement can be defined as being made up of two components: uncertainty derived from sampling a matrix and using those samples to represent the whole sampled mass; and the uncertainty derived from the analytical process [14, 15].

Now, the aim of this paper is to use advantages of fuzzy model to improve sampling and solve the problems that occur in probability sampling with the cross-disciplinary approaches foruncertainty analyses suitable for random number generation that is proposed by using fuzzy set theory and stratification sampling. In stratified and cluster sampling it is hard and time-consuming to separate data into categories, however after do this it is not precise. For example one element can be belonging to some categories near to that element (Figure 1).

As a result, although these methods have many benefits in sampling versus of other methods, unfortunately category based methods have many problems, because of unsuitable clustering results are not precise. A basic problem, at the present stage of the sampling methods, is how to manage the variance of sample process while taking into account its intrinsic features of uncertainty, including imprecision and vagueness. The proposed method is in fact a compatible approach to vary types of sampling methods, and appliedto improving the sampling efficiency in the uncertainties, so we decrease uncertainties to reach accuracy in results.

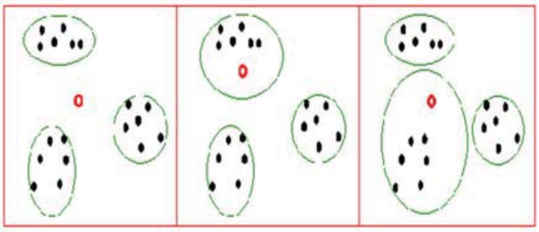

Figure 1. Methods of element categorizing in groups

The remainder of this paper is classified as follows: section-II presents the research background on quality of survey sampling results. Section-III presents a survey on stratified random sampling method. Section-IV provides Fuzzy sampling and the application of the Fuzzy set theory and Possibility extension theorem in generating the proposed Fuzzy-Strata sampling (FSS) algorithm. Finally, the last section presents the conclusions and discussion.

2. Quality of Survey Sampling Results

A survey is any activity that collects information in an organized and methodical manner about characteristics of interest from some or all units of a population using well-defined concepts, methods and procedures, and compiles such information into a useful summary form [16]. A survey usually begins with the need for information where no data – or insufficient data – exist. Researchers often use sample survey methodology to obtain information about a large population by selecting and measuring a sample from that population [17]. Due to variability among items, researchers apply scientific probability-based designs to select the sample [18]. When researchers describe the quality of survey results, they may use one or more of the following terms: Accuracy, Precision and Margin of error [16]. Accuracy refers to how close a sample statistic is to a population parameter, and precision refers to how close estimates from different samples are to each other. For example, the standard error is a measure of precision. When the standard error is small, estimates from different samples will be close in value; and vice versa. Precision is inversely related to standard error. When the standard error is small, sample estimates are more precise; when the standard error is large, sample estimates are less precise. Lastly, the margin of error expresses the maximum expected difference between the true population parameter and a sample estimate of that parameter. To be meaningful, the margin of error should be qualified by a probability statement. The margin of error is equal to half of the width of the confidence interval[19]. Once a choice is made to use a probability sample design, one must choose the type of probability sampling to use. There are four major types of probability sample designs: Simple Random Sampling, Stratified Sampling, Systematic Sampling, and Cluster Sampling. Simple random sampling is the most recognized probability sampling procedure. Stratified sampling offers significant improvement to simple random sampling. Systematic sampling is probably the easiest one to use, and cluster sampling is most practical for large national surveys [20].

3. Stratified Random Sampling

Random stratified sampling is one of the most well-known statistical methods to study the behavior of the average population of the studied variable. And has many applications in using real data sets and data needed for simulation [21], and also, able to estimate the distributions via Fuzzy logics [23], and multilevel Monte Carlo simulations [24]. In some cases, this sampling method has been used to study multivariate populations to estimate population characteristics [22]. A survey is any activity that collects information in an organized and methodical Stratified random sampling refers to a sampling method that has the following properties:

- The population consists of N elements.
- The population is divided into H groups, called strata.
- Each element of the population can be assigned to one, and only one, stratum.
- The number of observations within each stratum N_h is known, and $N = N_1 + N_2 + N_3 + ... + N_{H-1} + N_H$.

The researcher obtains a probability sample and draws a simple random sample from each stratum (**Figure 2**).

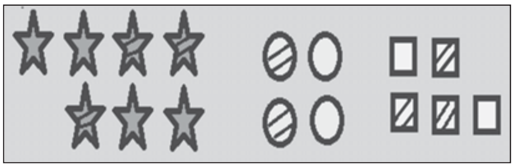

Figure 2. An example of stratified sampling

Stratified sampling offers several advantages over simple random sampling:

- A stratified sample can provide greater precision than a simple random sample of the same size.
- Because it provides greater precision, a stratified sample often requires a smaller sample, which saves money.
- A stratified sample can guard against an "unrepresentative" sample (e.g., an all-male sample from a mixed-gender population).
- We can ensure that we obtain sufficient sample points to support a separate analysis of any subgroup.

The main disadvantage of a stratified sample is that it may require more administrative effort than a simple random sample.

4. Fuzzy Sampling

In argued method each element just belongs to one category, but in our approach each element includes in all groups with one exception that membership values are different. For example, consider age ranges. (**Figure 3**) i.e. a man with 30 years old belongs to five groups by blow membership values:

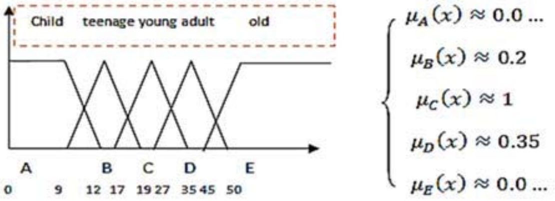

Figure 3. Demonstration of fuzzy membership function for age example

When such samples are categorized with stratified method, some errors raised for computing values like variance, mean, etc. in fuzzy approach these types of errors can be eliminated. In fact, each element is member of every category with a probability between 0 and 1. Our new approach is a combination of fuzzy and stratified. In this approach we assume that linguistic variables comprised as three parts and modeled in triangular. When we categories such samples with stratified method, some errors raised for computing values like variance, mean, etc. in fuzzy approach these types of errors can be eliminated or reduced.

4.1. Proposed FSS Algorithm

Stratified sampling offers several advantages over simple random sampling [1]: Like stratified method we choose some random elements from each stratum for sampling, and then we identify the neighborhood of strata to modeling fuzzy. Let n_1, n_2..., n_m be the number of elements in each of the m sampled strata. Let $\overline{X}_1$, $\overline{X}_2$,..., $\overline{X}_m$ be the means of the sampled strata. The relative uncertainty itself can also be used without the subsequent statistical testing particularly if enough is known about the parameter being evaluated. Variance reduction techniques are methods that attempt to reduce the variance, i.e., the dispersion associated with the variations, of the parameter being evaluated. This can result in one of two outcomes. Either the variance is reduced for the same number of sampling or the number of sampling can be reduced for the same variance, the comparison of both being made when no variance reduction techniques are used. This therefore either increases the confidence in the results or reduces the computational burden. There are many forms of variance reduction techniques and a specialized text should be consulted if full details of all techniques are needed. This algorithm provides reduced variances and standard errors comparison with other traditional methods. Schematic of proposed method is shown in **Figure 4**.

Steps that are used for programming with use of fuzzy-stratified method are mentioned in below algorithm (**Figure 4**), consist of following classified steps:

1. Select *m* random elements from each strata
2. Determine neighborhood of between strata
3. Calculate membership values for each strata
4. Calculate mean value, variance, max confidence and min confidence from formulas and
5. Complementary outputs of the algorithm are as follows:

a. Calculate values for each stratum with effect of strata neighbors

b. Calculate values with effect of element neighbors in each strata

c. Calculate values with effect of both element neighbors and strata neighbors

Formulations and relations that are used for computing the parameters such as Mean, Variance, Maximum and Minimum Confidence intervals are the same as stratified sampling formulations, in addition with fuzzy membership functions. Hence, for Mean estimator of this method we have:

To compute the overall sample mean, we need to compute the sample means for each stratum.

$$\overline{X}_i = \frac{\sum x_i}{n_i} \qquad (1)$$

Then;

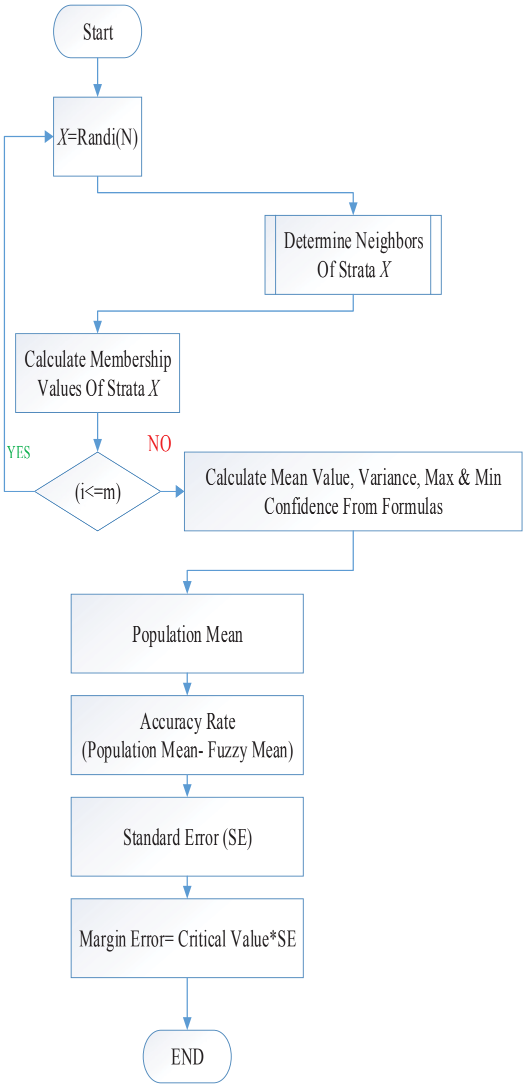
Start
X=Randi(N)
Determine Neighbors Of Strata X
Calculate Membership Values Of Strata X
YES
NO
(i<=m)
Calculate Mean Value, Variance, Max & Min Confidence From Formulas
Population Mean
Accuracy Rate (Population Mean- Fuzzy Mean)
Standard Error (SE)
Margin Error= Critical Value*SE
END

Figure 4. Flowchart of proposed sampling method

$$Mean = \frac{\sum_{i=1}^{m} (\overline{X}_i n_i + \mu_{i+1} \overline{X}_{i+1} n_{i+1} + \mu_{i-1} \overline{X}_{i-1} n_{i-1})}{\sum_{i=1}^{m} n_i} \qquad (2)$$

Where $\overline{X}_i$ and n_i are the mean and the number of elements for strata I respectively. μ_{i+1} and μ_{i-1} are membership values of two neighboring strata.

An estimator of samples variance is given by:

$$Variance = \frac{M-q}{Mq\overline{n}^2} \frac{\sum_{i=1}^{m} n_i^{\,2} (\overline{X}_i - Mean)}{q-1} \qquad (3)$$

Where

$$\overline{n} = \sum_{i=1}^{m} n_i \Big/ q \quad \text{And} \quad q = m(\mu_{i+1} + \mu_{i-1})$$

Standard error (SE):

$$SE = \left(\frac{1}{N}\right) \times \sqrt{\sum (N_i^2 \times \left(1 - \frac{n_i}{N_i}\right) \times Variance \Big/ n_i)} \qquad (4)$$

Where,

N And N_i are the number of observations in the population for strata I respectively.

Margin of error (ME):

$$ME = Critical \quad Value * SE = 1.96 * SE \qquad (5)$$

Specify the confidence interval. The range of the confidence interval is defined by the sample statistic ± margin of error. And the uncertainty is denoted by the confidence level, using the preceding information, we construct a 95% confidence interval for Mean as follows:

$$Confidence \quad Interval = Mean \pm 1.96 * Variance \qquad (6)$$

For example, at the end of every school year, the state administers a reading test to a sample of graders. The school system has 50,000 graders; first to fifth graders each has 10,000 students. This year, a proportionate stratified sample was used to select 500 students for testing. Because the population has equal students for every stratum, each stratum consisted of 100 students. We modeled this method, stratified sampling and exhaustive method (real mean and variance) in a system (win 7, mat lab R2010a, 2.30 GHz, 2 GB), and then compared results are given in **Table 11**.

Results show that significantly error values for fuzzy method is less than stratified and also approximations of mean value for fuzzy is close to population mean (Figure 5). Fuzzy method has equal differences SE like stratified method, as results in statistical approaches we can see that this method is an effective way to compute statistical values because of sufficient precision and ME. Standard Error deviations on difference run for fuzzy and stratified sampling are shown in Figure 6.

Also, Figure 7 shows the Mean values for exact, stratified sampling, and fuzzy sampling. As obviously, the fuzzy-stratified method has less deviation from the mean exact solution rather than simple stratified method. So, this method shows less error values from the exact solve, and we can conclude that accuracy rate for fuzzy method is higher than stratified.

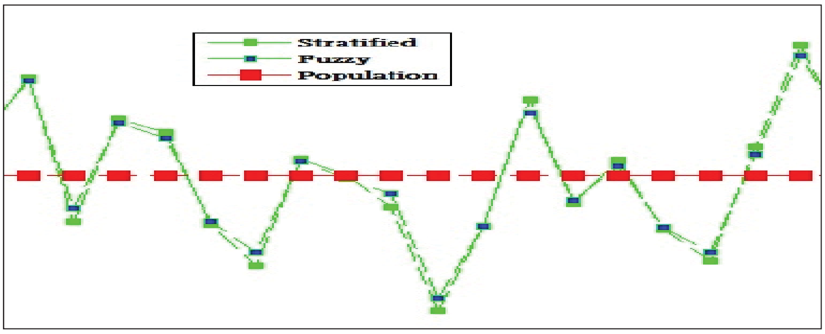

Figure 5. Mean improving using fuzzy sampling

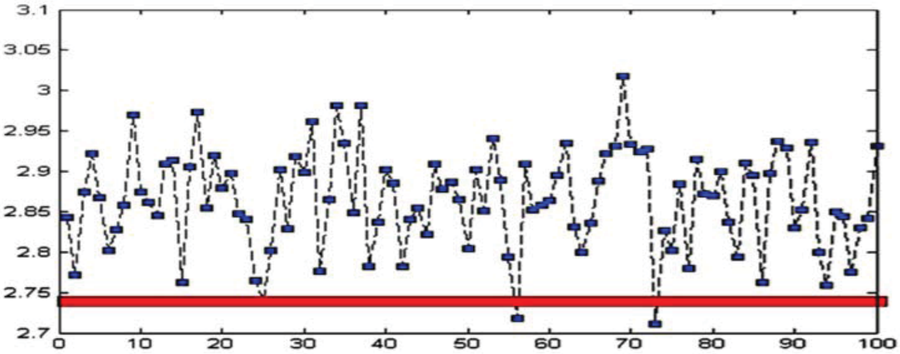

Figure 6. Standard Error deviations on difference run for fuzzy and stratified sampling

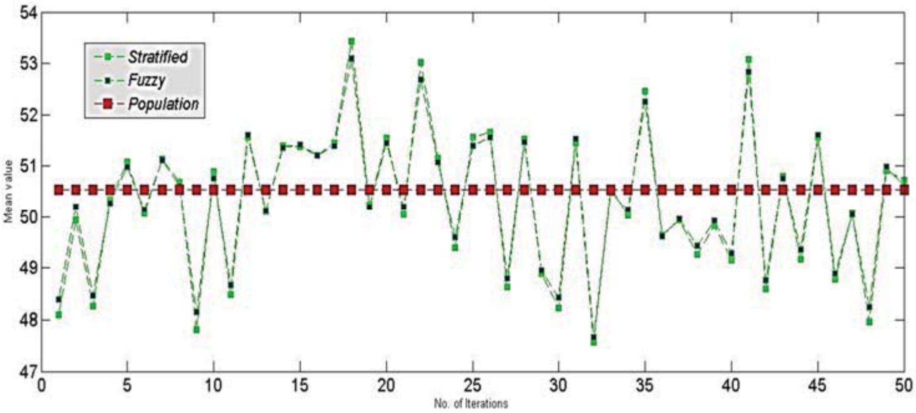

Figure 7. Mean deviations of sampling

Table 11. Sampling Parameter Results

Method	Mean Value	Standard Error	Confidence interval	Margin Error	Accuracy (%)	No. of Iterations
Population (Exhaustive)	50.5445	2.739	[45.1759, 55.913]	5.3685	100	100
	"	"	"	"	"	1000
	"	"	"	"	"	10000
Stratified	50.3026	2.8645	[44.6881, 55.9171]	5.6145	97.721	100
	50.5346	2.8714	[44.9066, 56.1627]	5.628	97.963	1000
	50.5623	2.8695	[44.9381, 56.1865]	5.6242	97.973	10000
Fuzzy	50.3382	2.8647	[44.689, 55.9186]	5.6148	97.938	100
	50.5381	2.8716	[44.9072, 56.1639]	5.6284	98.161	1000
	50.5595	2.8697	[44.9389, 56.188]	5.6246	98.15	10000

10000

Fuzzy	50.3382	2.8647	[44.689, 55.9186]	5.6148	97.938	100
	50.5381	2.8716	[44.9072, 56.1639]	5.6284	98.161	1000
	50.5595	2.8697	[44.9389, 56.188]	5.6246	98.15	10000

As shown in Figure 7, the purpose of this study is to show that the mean values in stratified sampling have a higher dispersion than the mean values in the fuzzy state. The values shown by the fuzzy diagram are in fact the values of the proposed fuzzy-strata method, and the strata diagram is the crude state of the strata sampling. It is clear that the deviations from the mean line in fuzzy-strata sampling is less. In Table-1 (7th column) as accuracy percentage shows the superiority of proposed method, as can be seen, the accuracy of fuzzy-strata for example for 1000 iteration is 98.161 is higher than the accuracy of the crude strata sampling (97.963). And for other number of iterations also is obvious.

V- Conclusion and Discussion

With the set of given numbers, elements or objects in the statistical approaches we can select the elements in a random way to compute values of many subjects like mean, variance and etc. There are some sampling methods in this way that argued in many publications. In the whole of these methods we assumed that the input data is precise, i.e. each data belongs to a set. For example, in stratified sampling we assume that the population of *N* units may be divided into m groups and the m strata are no overlapping, and then samples select from within each groups. However, in real world, examples and elements may be belonging to any groups or not precise. Even in simple random sampling or systematic way we can see some errors lead to equal probabilities of elements. In this paper we introduce new approach base on fuzzy theory combined with sampling methods. The process is to calculate the relative error or uncertainty as the simulation proceeds and, using appropriate statistical tests generally based on the fuzzy theory, calculate the confidence interval after each sampling. This is compared with the pre-specified example. The case study results show that the proposedmethod provides better measure of uncertainty than the existing methods as unlike traditional samplingmethod. In this paper, in order to decrease the uncertainty of sampling, the variance reduction techniques have been used. Variance reduction techniques are methods that attempt to reduce the variance, i.e., the dispersion associated with the variations, of the parameter being evaluated.

References

[1] M. Nadjafi, M.A. Farsi, and A. Najafi, Uncertainty improving in importance sampling: An integrated approach with Fuzzy-Cluster sampling, 24th annual European Safety and Reliability (ESREL) Conference, Wroclaw, Poland, 14-18 September, 2014.

[2]. Kurowicka, D. and R.M. Cooke, Uncertainty analysis with high dimensional dependence modelling. 2006: John Wiley & Sons.

[3]. Ondrej Linda and M. Manic, 'Importance Sampling Based Defuzzification for General Type-2 Fuzzy Sets',WCCI 2010 IEEE World Congress on Computational Intelligence, July, 18-23, 2010 - CCIB, Barcelona, Spain.

[4]. Verdonck, F.A., et al., Improving uncertainty analysis in European Union risk assessment of chemicals. Integrated environmental assessment and management, 2007. 3(3): p. 333-343.

[5]. Kwakernaak, H., Fuzzy random variables—I. Definitions and theorems. Information Sciences, 1978. 15(1): p. 1-29.

[6]. L. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets Syst. 1 (1978) 3-28.

[7]. Viertl, R., Statistical methods for fuzzy data. 2011: Wiley. com.

[8]. Zadeh, L.A., The concept of a linguistic variable and its application to approximate reasoning—I. Information sciences, 1975. 8(3): p. 199-249.

[9]. Garibaldi, J.M. and R.I. John. Choosing membership functions of linguistic terms. in Fuzzy Systems, 2003. FUZZ'03. The 12th IEEE International Conference on. 2003. IEEE.

[10].Bethlehem, J., The rise of survey sampling. CBS Discussion Paper, 2009. 9015.

[11].Institute, S., SAS/STAT 12. 1 User's Guide: Survey Data Analysis (Book Excerpt). 2012: SAS Institute.

[12].Levy, P.S. and S. Lemeshow, Sampling of populations: methods and applications. 2013: John Wiley & Sons.

[13].Hansen, M.H. and W.N. Hurwitz, Sample survey methods and theory. Vol. I. 1953.

[14].Grøn, C., et al., Uncertainty from sampling–A Nordtest handbook for sampling planners on sampling quality assurance and uncertainty estimation. Nordtest Report TR, 2007. 604.

[15].Billinton, R. and R.N. Allan, Reliability evaluation of engineering systems: concepts and techniques. 1983: Plenum Press New York, NY.

[16].Molenberghs, G., Survey methods and sampling techniques. Limburg, Belgium: Center for Statistics, Universiteit Hasselt, 2008.

[17].Lyberg, L.E., et al., Survey measurement and process quality. Vol. 999. 2012: John Wiley & Sons.

[18].Lohr, S., Sampling: design and analysis. 2009: Cengage Learning.

[19].Chaudhuri, A. and H. Stenger, Survey sampling: theory and methods. 2010: CRC Press.

[20].Cochran, W.G., Sampling techniques. 2007: John Wiley & Sons.

[21]. Ghufran, Shazia, Srikant Gupta, and Aquil Ahmed. "A fuzzy compromise approach for solving multi-objective stratified sampling design." Neural Computing and Applications 33.17 (2021): 10829-10840.

[22]. Mradula, et al. "Efficient estimation of population mean under stratified random sampling with linear cost function." Communications in Statistics-Simulation and Computation (2019): 1-24.

[23]. Haq, Ahteshamul, Irfan Ali, and Rahul Varshney. "Compromise allocation problem in multivariate stratified sampling with flexible fuzzy goals." Journal of Statistical Computation and Simulation 90.9 (2020): 1557-1569.

[24]. Taverniers, Søren, and Daniel M. Tartakovsky. "Estimation of distributions via multilevel Monte Carlo with stratified sampling." Journal of Computational Physics 419 (2020): 109572.