A Simulation Approach to Studying Normality of Sampling Distribution

***Ibrahim, S.A.**

Department of Physical Sciences, Al-Hikmah University, Ilorin, Nigeria

Received: February 23, 2015; Revised: February 11, 2016; Accepted: March 29, 2016

Abstract

The normality of sampling distribution is crucial for statistical inference; and sampling distribution is the source of all knowledge in statistical analysis. In this article, sampling distributions of means for normal and uniform distributions were studied using a simulation approach. Data used were simulated from R software. It was found that the mean of the sampling distribution of means for all sample size n considered for each distribution were in a neighborhood of the true population mean with a little bias. It was also found that the standard deviation (also called standard error) of sampling distribution of means agree quite closely with the formula based estimate, and standard error decreased as sample size n increased for both distributions considered. Finally, it was observed that the histogram visualized the meaning of Central Limit Theory and the shape of the sampling distribution of means approximate normal for each sample size n and for the population distributions considered. Since the sampling distribution of means is approximately normal, this justifies the use of statistical procedure based on normal distribution theory to estimate confidence intervals of μ even when working with non-normal data.

Keywords: Normal Distribution, Sampling Distribution, Central Limit Theorem, Simulation.

1.0 Introduction

The normality of the sampling distribution is secured if the parent population from which samples are drawn has a normal distribution. Hence, the normality of sampling distribution (i.e. normality assumption) is crucial for statistical inference [1]. Sampling distribution (SD) and some other basic statistical concepts such as Central Limit Theorem (CLT), Standard Error (SE), Confidence Intervals (CI), etc. are critical in gaining necessary statistical skills but are difficult to understand due to the abstract nature of the concepts. Computers have changed the face of Statistics due to its fast speed and flawless accuracy with respect to large data set acquired by researchers, which made them indispensable for many modern analyses. Thus, computer simulation method is an active learning approach which uses physical activities to demonstrate abstract concepts. The method allows researchers experiment with data and visualize the result of sampling distribution easily; and if the histogram of the sampling distribution of means conforms quite well to the limiting normal distribution for all sample means, the confidence interval estimates of μ should therefore be based on that distribution [2, 3].

Researchers have used different approaches to explore sampling distribution including development of computer simulation software (or program) to provide active learning of the sampling distribution and CLT as well as to explore the properties of these concepts [4]. Also, the Excel data table was demonstrated to study the sampling distribution through simulation of random sampling from a specified population to generate sampling distribution of means and to understand the central limit theorem in introductory statistics class [5]. A dynamic statistics environment (Fathom) was used to understand the relationships among sample size *n*, and the behavior of sampling distributions [6].

A comprehensive literature review has been done on computer simulation method used in major areas of abstract concepts in statistics [7]. However, none of the researchers' considered have used R software to explore the properties of the sampling distribution and this is an obvious gap in the literature. R software is known as a new develops software that promotes graphic and numerical visual display. Also, R is a high-level language and an environment for data analysis and graphics. A large proportion of the world's leading statisticians use R for data analysis. More people and researchers are reporting their results in the context of R [8, 9].

*** Corresponding Author: Tel: +234(0)8052262175; Email: adeshinas2010@alhikmah.edu.ng, adeshinas@gmail.com**

In the real world, a sample of size n was usually observed, and therefore gives just one estimate. A Monte Carlo simulation experiment allows replicating the real world study many (R) times. Each time, a different sample of size n is drawn from the original population. Thus an estimate of each sample is calculated and any estimate will be a bit different. The empirical distribution of these many estimates approximates the true estimator. Many statistical techniques and simulation methods for sampling distribution have been described [4,5,7,10]. Therefore, this paper is aimed at studying the normality of the sampling distribution of means in normal and uniform distributions with particular interest in the properties of sampling distribution and to explore fundamental of studying the concept using R software.

2.0 Materials and Methods

2.1 Materials

Data (also called materials) used for this study were obtained through computer simulation with R console version 2.8.1 together with Tinn-R version 2.6.2. Thus, two separate sampling experiments are considered for this study.

1. Sampling experiment A takes 10,000 repeated random samples of size *n* = [10, 20, 30, 50, 80, 100] from a normal distribution each time, with mean $\mu = 6.0$ and standard deviation $\sigma = 3.29$ through R code. The mean of each repeated sample was calculated and the means are arranged to form a distribution **i**n order to demonstrate the hypothetical sampling distribution of means as well as to verify the validity of its properties under normal distribution.

2. Sampling experiment B takes 10,000 repeated random samples of size *n* = [10, 20, 30, 50, 80, 100] from a uniform distribution (0,100) each time, with mean $\mu = 50$ and standard deviation $\sigma = 28.87$ through R code. The mean of each repeated sample was calculated and the means are arranged to form a distribution in order to demonstrate the hypothetical sampling distribution of means as well as to verify the validity of its properties under the uniform distribution.

2.2 Methods

2.2.1 Sampling Distribution of the Mean

The gateway to statistical inference is the sampling distribution. The Sampling distribution of means describes the probability distribution of sample mean based on all possible simple random samples of the same size *n* from the same population. The properties of the sampling distribution of the means are stated as follows:

- Sample means are random
- The mean of all sample means is equal to the population mean $\mu_{\hat{\theta}} = \mu$.
- The standard deviation of the sample means (also called standard error) is equal to the population standard deviation divided by the square root of the sample size $\sigma_{\hat{\theta}} = \frac{\sigma}{\sqrt{n}}$.
- The sampling distribution of sample means will closely approximate the normal distribution as sample size *n* increases.

2.2.2 General Simulation Procedures

Step 1: Generate independent draws, large sample of size *n*, from a population distribution of interest

Step 2: For each sample, compute the statistic of interest says sample mean, and denote it by $\hat{\theta}$

Step 3: Repeat step 1 and 2 R times, hence, the following estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, ..., \hat{\theta}_R$ are obtained.

Step 4: Construct a relative frequency histogram from R number of estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, ..., \hat{\theta}_R$. The distribution obtained is called sampling distribution of the estimator $\hat{\theta}$. Thus the distribution can be used to make inference about the parameter θ. Also, based on the distribution obtain in step 4, mean of sampling distribution of means (MSDM), bias, standard error (SE) and 95% confidence interval (CI) are stated as follows:

$$\text{MSDM} \; \mu_{\hat{\theta}} = \frac{1}{R}\sum_{i=1}^{R} \hat{\theta}_i$$

$$Bi\hat{a}s = \mu_{\hat{\theta}} - \mu$$

$$S.E\sigma_{\hat{\theta}} = \sqrt{\frac{1}{R-1}\sum_{i=1}^{R}(\hat{\theta}_i - \frac{1}{R}\sum_{i=1}^{R}\hat{\theta}_i)^2}$$

$$CI_{.95} = \mu_{\hat{\theta}} \pm z_{.025} * S.E\sigma_{\hat{\theta}}$$

2.2.3 *The Central Limit Theorem (CLT)*

CLT explains the shape of the sampling distribution. This theorem state that for a population of any distribution, the distribution of the sample mean approaches a normal distribution as the sample size *n* increases, the larger the sample size *n*, the better the approximation. On the basis of this theorem, normal distribution could be used for statistical inference about the mean for large sample sizes, even if the original population is not normally distributed. Computer simulation would aid visualizing this important abstract concept which many people have been used without understanding how the underlying concepts work.

2.2.4 *Normal Distribution*

An important property of the normal random variable is that it approximates the distribution of the average of a sample taken from any distribution and this remarkable result is based on the CLT [11]. The Normal distribution is a bell-shaped curve which extends indefinitely in both directions. It is symmetrical round the mean of the variable, whose values are measured in the horizontal axis. The vertical axis depicts the value of probability density function $f(x_i)$. The equation of the normal curve is $f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2$

Where $f(x_i)$ = probability of x assuming the value x_i

μ = the mean of variable x_i

σ = the standard deviation of x

π = 3.14159

exp = 2.71828, the base of natural logarithms

2.2.5 *Characteristics of Normal Distribution*

The distribution is symmetric and illustrated as a bell-shaped curve. Two parameters mean μ and standard deviation σ, completely determine the location and shape of a normal distribution. The highest point on the normal curve is at the mean, which is also equal to the median and mode. The mean can be any numerical value: negative, zero, or positive. The total area under the curve is one (0.5 to the left of the mean and 0.5 to the right). The area under the normal curve represents the probability of a normal random variable. Thus 68.27% of cases will lie within 1 standard deviation of the mean, 95.45% within 2 standard deviations, and 99.73% within 3 standard deviations.

2.2.6 Histogram

Histogram is a graphical representation of the distribution of data. It is an estimate of the probability distribution of a continuous variable (quantitative variable). A histogram may also be normalized displaying relative frequencies. It then shows the proportion of data that fall into each of several categories, with the sum of the heights equaling 1. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent and usually of equal size. The rectangles of a histogram are drawn so that they touch each other to indicate that the original variable is continuous.

2.27 Evaluation of Normality of Sampling Distribution of Means

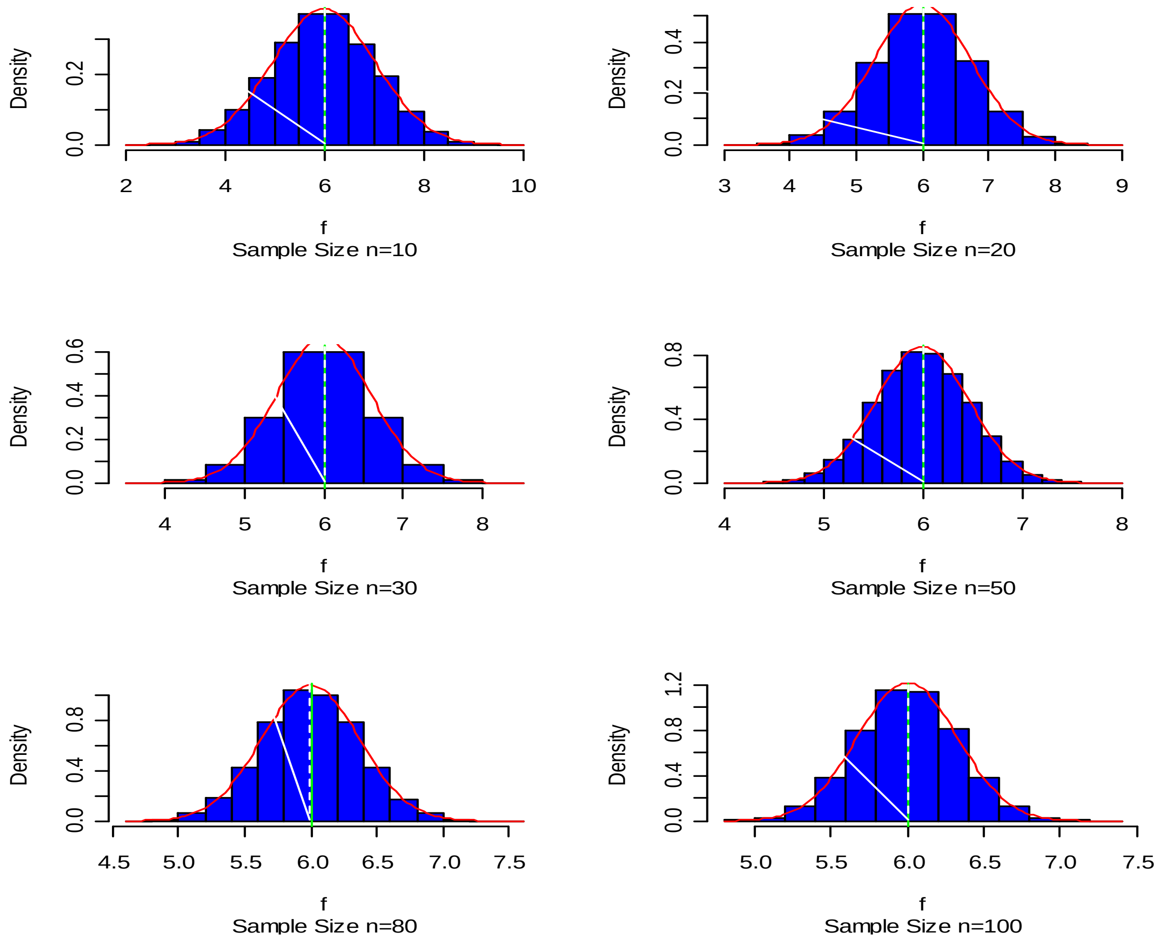
One simple way to judge the normality of the sampling distribution of means in R code is to get R code to superimpose a normal curve on the histogram of the actual sampling distribution of means of a repeated sampling gathered through simulation. R code provides the version of the normal curve that fits the mean and standard deviation of a data set. The researcher then compared normal curve with the histogram of sampling distribution of means as regard to positive and negative skewness or normally distributed in order to make a valid conclusion.

3.0 Results

R code is used to simulate 10,000 repeated random samples of size *n* = [10, 20, 30, 50, 80, 100] from normal distribution $(\mu = 6.0, \sigma = 3.29)$. Thus, the summary statistics obtained is as presented in Table 1 while the distribution of means at different sample sizes is presented in Fig. 1. Also, R code is used to simulate 10,000 repeated random samples of size *n* = [10, 20, 30, 50, 80, 100] from a uniform distribution (0,100) with (μ =50, σ =28.5). Thus, the summary statistic obtained is presented in Table 2 followed by distribution of means at different sample sizes in Fig. 2. The line plot showing the relationship between estimates of standard error obtained from normal and uniform distributions at different sample size *n* is presented in Fig. 3.

Table 1: The summary statistic of 10,000 sampling distribution of means for $N(6.0,\ 3.29)$ at different sample size *n*

Sample Size *n*	**MSDM** $\mu_{\hat{\theta}}$	**Standard Deviation (SE)** $\sigma_{\hat{\theta}}$	**Bias=** $\mu_{\hat{\theta}} - \mu$	**Formula S.E=** $\frac{\sigma}{\sqrt{n}}$	**95% Confidence Interval (CI)**
10	5.992617	1.037576	-0.007383	3.29/ $\sqrt{10}$ =1.04039	3.958968, 8.026266
20	5.998002	0.7291502	-0.001998	3.29/ $\sqrt{20}$ =0.7356	4.568868, 7.427136
30	5.998614	0.5976099	-0.001386	3.29/ $\sqrt{30}$ =0.60067	4.827299, 7.16993
50	5.998409	0.4671169	-0.001591	3.29/ $\sqrt{50}$ =0.46528	5.08286, 6.913958
80	5.997869	0.367946	-0.002131	3.29/ $\sqrt{80}$ =0.36783	5.276695, 6.719043
100	6.000154	0.3278165	0.000154	3.29/ $\sqrt{100}$ = 0.329	5.357634, 6.642674



Sample Size n=10

Sample Size n=20

Sample Size n=30

Sample Size n=50

Sample Size n=80

Sample Size n=100

Fig. 1: Distribution of 10,000 sampling distribution of means (obtained from normal distribution) at different sample size *n* = [10, 20, 30, 50, 80, 100]

Table 2: The Summary statistics of 10,000 sampling distribution of means for uniform population distribution $U(0,100)$ with μ =50, σ =28.87 at different sample size *n*.

Sample Size *n*	**MSDM $\mu_{\hat{\theta}}$**	**Standard Deviation (SE) $\sigma_{\hat{\theta}}$**	**Bias = $\mu_{\hat{\theta}} - \mu$**	**Formula SE=$\frac{\sigma}{\sqrt{n}}$**	**95% Confidence Interval (CI)**
10	49.96247	9.144485	-0.03753448	$28.87/\sqrt{10}$ = 9.1295	32.03928, 67.88566
20	49.98531	6.39048	-0.01469192	$28.87/\sqrt{20}$ =6.455528	37.45997, 62.51065
30	50.01001	5.237782	0.01001069	$28.87/\sqrt{30}$ =5.270917	39.74396, 60.27606
50	49.97174	4.063881	-0.02826474	$28.87/\sqrt{50}$ =4.082835	42.00653, 57.93695
80	49.98958	3.216327	-0.01042170	$28.87/\sqrt{80}$ =3.227764	43.68558, 56.29358
100	49.99223	2.899121	-0.007772398	$28.87/\sqrt{100}$ = 2.887	44.30995, 55.67451

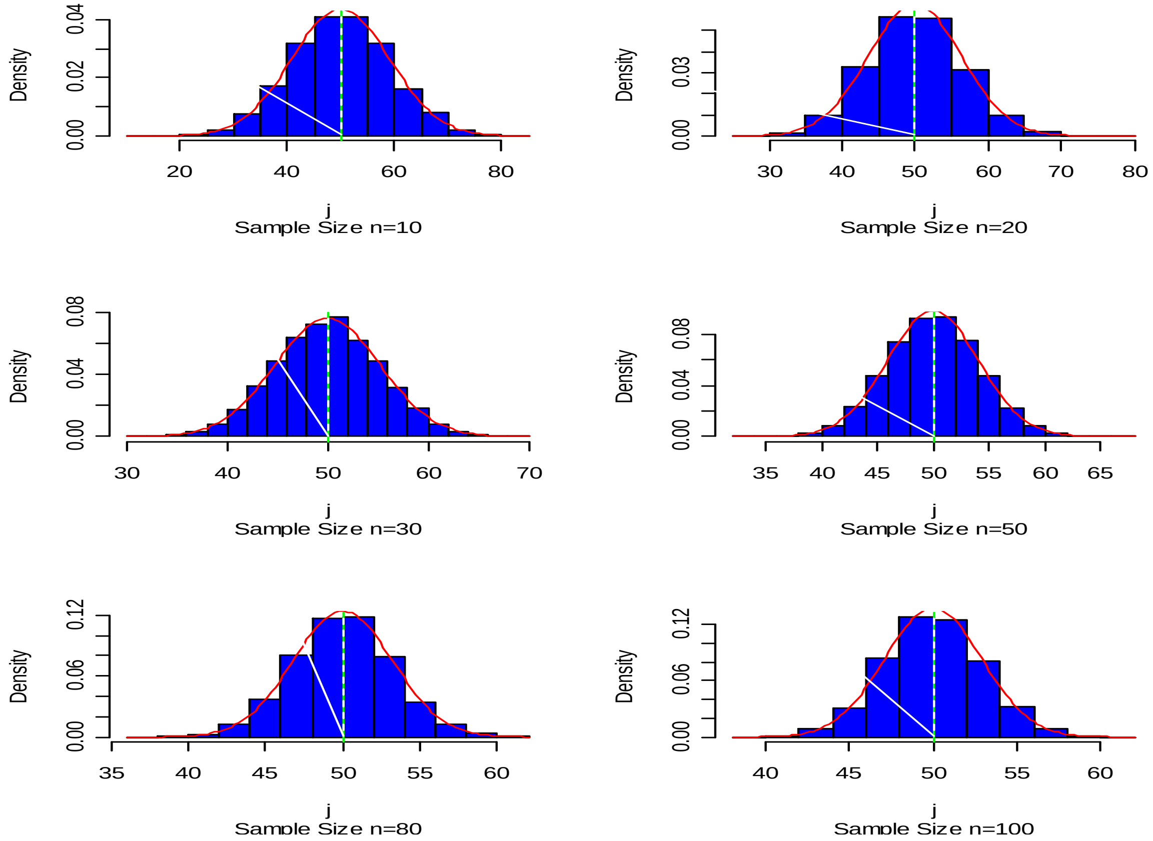


Fig. 2: Distribution of 10,000 sampling distribution of means (obtained from uniform distribution) at different sample size *n* = [10, 20, 30, 50, 80, 100].

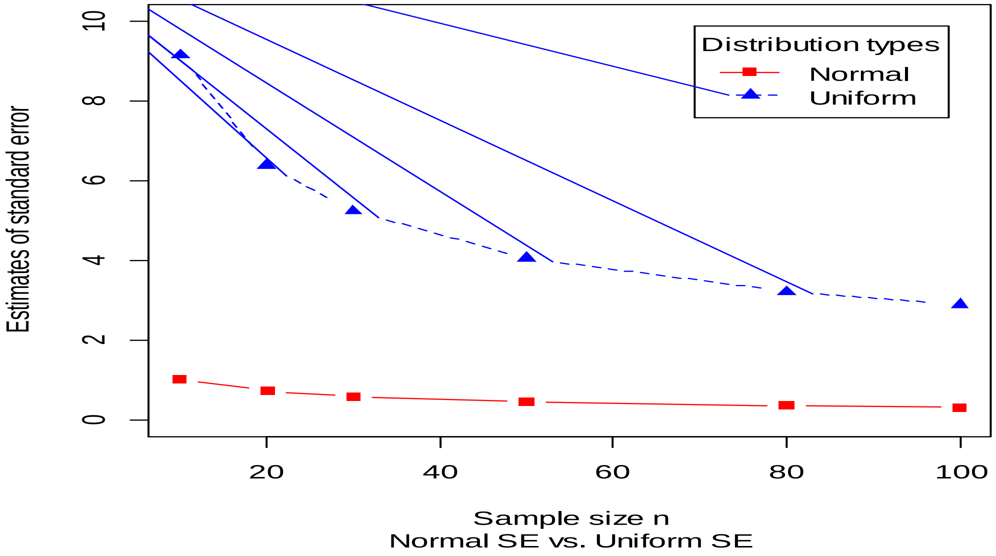


Fig 3: Line plot showing the relationship between estimates of standard error obtained from normal and uniform distributions at different sample size *n*.

4.0 Discussion

In this study, 10,000 repeated random samples of size *n* = [10, 20, 30, 50, 80,100] are generated each time from normal and uniform distributions, through R code, and the mean of each repeated random sample is computed and the means are arranged to form a distribution in order to demonstrate the hypothetical sampling distribution of means, verify the validity of its properties as well as to make statistical inference on it. It was found that sample means, for each sample, were not very spread out but random for both normal and uniform distributions. All were in a neighborhood of the true population mean. As the sample size *n* increases, the mean of the sampling distribution of means $\mu_{\hat{\theta}}$ gets closer and closer to the population mean μ. For instance, the actual population value of the normal distribution $\mu = 6.0$ is attained when sample size *n* was 100 while uniform distribution actual population value $\mu = 50$ is attained when sample size *n* was 30. In this study, the mean of the sampling distribution of means has a little bias as an estimator of the population mean considered. This little difference indicates that sample mean is roughly an unbiased estimate for the true population mean considered as confirmed from Table 1 and Table 2.

Standard deviation of sampling distribution of means also called standard error of means (SE) $\sigma_{\hat{\theta}}$ is used to measure accuracy of the mean of the sampling distribution of means $\mu_{\hat{\theta}}$. The standard deviation of means $\sigma_{\hat{\theta}}$ is less than the population standard deviation σ for each sample size n and for each population distribution considered. Also, SE decreases as the sample size *n* increases i.e. the sampling error in estimating μ decreases when sample size *n* increases. In terms of comparison, it was observed that the estimates of standard error obtained from a normal distribution approached zero faster than uniform distribution at different sample size *n*. however, this observation was based on the choice of parameter of the distribution used as confirmed from Fig. 3.

Histogram allowed for visualizing the meaning of CLT and the effect of sample size *n*. The shape of the sampling distribution of means approximately normal for each sample size n and for population distributions considered. Since the sampling distribution of average $\hat{\theta}$ is approximately normal ie. $N(\mu, \sigma^2/n)$, this justified use of statistical procedure based on normal distribution theory for construction of interval estimation that contains population μ. As a result of this, the standard normal distribution probabilities table is used in this study to compute 95% confidence interval for μ for each sample size *n* as a statistical inference, for each distribution considered as confirmed in Table 1 and Table 2. It was also found that averaging over many observations is more accurate than just looking at one or two observations. The results obtained in this study agree with earlier reports where the distribution of the sample means look very close to a normal distribution as sample size *n* increases [2-5, 12].

5.0 Conclusion

Computer simulation has aided visualizing the important abstract concepts namely normality, sampling distribution and central limit theorem which many researchers have used without understanding how the underlying concepts work. Sampling distributions of means are approximately normal regardless of the shape of the parent population (normal or non-normal). As a result of this, the standard normal distribution probabilities table is used to compute 95% confidence interval for μ.

6.0 Acknowledgement

The contributions of reviewers, Editor and my colleagues in this work is acknowledged and highly appreciated.

References

[1] Koutsoyiannis, A. (1977). Theory of Econometrics. 2^{nd} edn. New York: Palgrave, pp.554.

[2] Carolyn, J A. (2005). Sampling Distributions, the CLT and Estimation. Ed Psych 580, pp. 45-46

[3] Tim, H., Shaun, M., David, S. M., Ashley, C., and Rachel, E. (2003). The Practice of Business Statistics: Bootstrap Methods and Permutation Tests. New York: W. H. Freeman and Company, pp. 18-22.

[4] Robert, C. d., Joan, G., & Beth, L. C. (1999). A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning. Journal of Statistics Education, Vol. 7, No. 3, pp. 3–7.

[5] Chandrakantha, L. (2014). Excel Simulation as a Too in Teaching Sampling Distributions in Introductory Statistical. In: Makar, K., De Sousa, B. and Gould, R. (Eds.). Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS). New York, USA: International Statistical Institute, pp.1-3.

[6] Ernesto, S. and Santiago, I. (2006). Meanings' Construction About Sampling Distribution in a Dynamic Statistics Environment. ICOTS-7, pp.1–4.

[7] Mills, J. D. (2002). Using Computer Simulation Methods to Teach Statistics: A Review of the Literature. Journal of Statistics Education, Vol. 10, No. 1, pp. 5-12.

[8] Crawley, M. J. (2007). The R Book. John Wiley & Sons Ltd, England.

[9] Robert, I. K. (2011). R in action: Data Analysis and Graphic with R. Manning Publications Co., Shelter Island.

[10] David, P. D. (2004). Using Simulation to Teach Distribution. Journal of Statistics Education, Vol. 12, No.1. pp. 1–3.

[11] Hamdy, A.T. (1997). Operations Research: An Introduction, 8th ed. United States of America: Pearson Education, Inc, pp.478-479

[12] Suat, S. and Dervis, T. (2007). Bootstrap and Jackknife Resampling Algorithms for Estimation of Regression Parameters. Journal of Applied Quantitative Methods, Vol. 2, No. 2, pp. 196–198.