*Transactions of the Nigerian Association of Mathematical Physics*
*Volume 12, (July – Sept., 2020 Issue*), **pp63 – 68**


# VISUALIZATION OF SOME ABSTRACT CONCEPTS IN STATISTICS

*S. A. Ibrahim and Y. S. Gorah*

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

***Abstract***

***In this study, visualization of some important abstract concepts in statistics were studied through simulation for three different probability distributions. The abstract concepts understudied includes: sampling distribution of means, normal curve, central limit theorem, standard deviation and standard error. Data used was generated through R statistical software. The results showed that sample means, for each sample, were not very spread out but random for each distribution. All were in a neighborhood of the true population mean.***
***Standard error of means decreases as the sample size n increases. Finally, the shape of sampling distributions of means is approximately normal regardless of the shape of the parent population (normal or non-normal). As a result of this, statistical procedure based on normal distribution theory is used to compute confidence interval for population mean, μ.***

***Keywords*:** Sampling Distribution of Mean, standard error, Normal Curve, Probability Distribution

## 1.0 Introduction

Concepts such as sampling distribution of means (SDM), normal distribution (ND), central limit theorem (CLT), Standard deviation (SD), standard error (SE) etc. are critical in gaining necessary statistical skill, in inferential statistics such as hypotheses testing and confidence interval. However, these concepts are difficult to understand due to abstract nature of them. Meanwhile, with the aid of computer software application such as R, these abstract concepts could be made concrete form easily, for details see [1-5]. This study is aimed at studying sampling distribution of means in some selected probability distribution; verify the validity of its properties; visualizing the important abstract concepts as well as to justified the use of statistical procedure based on the normal distribution theory. The probability distributions selected for the study includes: Normal, Chi-square, and Laplace distributions respectively. The following books were used in writing the R codes for this study [6, 7].

## 2.0 Materials and Methods

Mathematical equations of some selected probability distributions used in this study were described as follow:

### 2.1 Normal Distribution Function

The probability density function of the normal distribution is:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2} \quad (1)$$

Where $\pi = 3.141559$ and $e = 2.71828$

#### 2.1.1 Properties of normal distribution

$$\text{mean} = \mu \quad (2)$$

$$\text{variance} = \sigma^2 \quad (3)$$

$$\text{standard deviation} = \sqrt{\sigma^2} \quad (4)$$

### 2.2 Chi-square Distribution Function

The probability density function of the chi-square distribution is given by:

$$f(x) = \frac{1}{2^{\frac{v}{2}}} x^{\frac{v}{2}-1} e^{-x/2} \quad (5)$$

Correspondence Author: Ibrahim S.A., Email: Adeshina2010@alhikmah.edu.ng, Tel: +2348052262175

**2.2.1 Properties of chi-square distribution**

Mean $= v$ (6)

Variance $= 2v$ (7)

Standard deviation $= \sqrt{2v}$ (8)

**2.3 Laplace Distribution Function**

The probability density function of the Laplace distribution is:

$f(x) = \frac{1}{2b} e^{-|x-\mu|/b}$ (9)

**2.3.1 Properties of Geometric Distribution**

Mean $= \mu$ , (10)

Variance $= 2b^2$ (11)

Standard deviation = $\sqrt{2b^2}$ (12)

**2.4 General Simulation Procedures used in this study**

Step 1: Generate independent draws, large sample of size *n*, from a population distribution of interest, such as normal, chi-squares distribution etc.

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 $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_R$ are obtained.

Step 4: Construct a relative frequency histogram from $R$ number of estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_R$. The distribution obtained is called sampling distribution of the estimator, $\hat{\theta}$. Thus the distribution can be used to make inference about the parameter, $\theta$.

Step 5: 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:

MSDM $\mu_{\hat{\theta}} = \frac{1}{R}\sum_{i=1}^{R} \hat{\theta}_i$ (13)

Bias = $\mu_{\hat{\theta}} - \mu$ (14)

$S.E\sigma_{\hat{\theta}} = \sqrt{\frac{1}{R-1}\sum_{i=1}^{R}(\hat{\theta}_i - \frac{1}{R}\sum_{i=1}^{R}\hat{\theta}_i)^2}$ (15)

$CI_{.95} = \mu_{\hat{\theta}} \pm Z_{.025} * S.E\sigma_{\hat{\theta}}$ (16)

**2.5 Simulation design**

Following Ibrahim [1] a simulation study was carried out, whereby three separate sampling experiments are considered for this study.

1. Sampling experiment A takes 1,000 repeated random samples of size *n* = (25, 35, 65, 85, 105, 135) from a normal distribution each time, with mean μ = 7.0and standard deviation $\sigma$ = 2.29 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. 1. Sampling experiment B takes 1,000 repeated random samples of size *n* = (25, 35, 65, 85, 105, 135) from a Chi-square distribution each time, with mean μ = v = 2 and standard deviation $\sigma = \sqrt{2v} = 2$ 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 Chi-square distribution.
3. 1. Sampling experiment C takes 1,000 repeated random samples of size *n* = (25, 35, 65, 85, 105, 135) from a Laplace distribution each time, with mean μ = 4 and standard deviation $\sigma = \sqrt{2b^2} =$ 1.4142 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 Laplace distribution.

**Table 1: Some selected probability distributions together with parameter specified arbitrary**

| **Distribution** | **Parameters** | **Mean μ** | **Std $\sigma$** |
|---|---|---|---|
| Normal | μ = 7.0 , $\sigma$ = 2.29 | μ = 7.0 | $\sigma$ = 2.29 |
| Chi-square | v =2 | μ = v = 2 | $\sigma = \sqrt{2v} = 2$ |
| Laplace | μ = 4 , $\sigma$ = 1 | μ = 4 | $\sigma = \sqrt{2b^2} =$ 1.4142 |

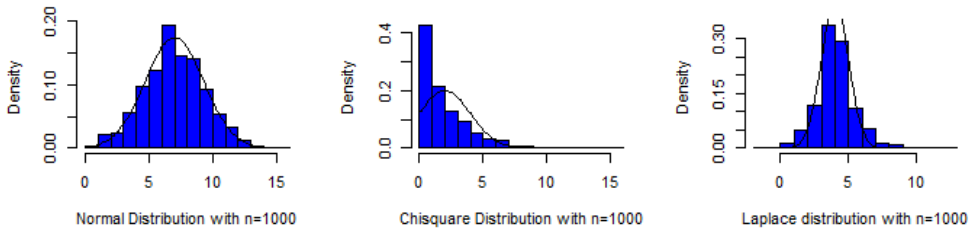


**Figure 1: Graphical presentation** of some selected probability distributions used in the study

### 3.0 Simulation Results

Summary statistic obtained under different probability distributions at different sample sizes is as presented in Table 2 through Table 4. Also, distribution of sampling distribution of means at different sample sizes was presented in Figure 2 through Figure 4.

**Table 2:** The summary statistic of 1,000 sampling distribution of means for normal distribution with mean, $\mu = 7.0$ and standard, $\sigma = 2.29$, at different sample size,$n$

| Sample size $n$ | MSDM =$\mu_{\bar{x}}$ | Standard deviation = $\sigma_{\bar{x}}$ | Bias = $\mu_{\bar{x}}$ - $\mu$ | Formula S.E $\frac{\sigma}{\sqrt{n}}$ | 95% confidence interval |
|---|---|---|---|---|---|
| 25 | 7.0021 | 0.4284 | 0.0021 | 0.458 | 6.1625, 7.8418 |
| 35 | 7.0027 | 0.371 | 0.0027 | 0.3871 | 6.2754, 7.7299 |
| 65 | 6.9919 | 0.2856 | -0.0081 | 0.284 | 6.432, 7.5517 |
| 85 | 6.9945 | 0.2444 | -0.0055 | 0.2484 | 6.5155, 7.4735 |
| 105 | 6.993 | 0.2193 | -0.007 | 0.2235 | 6.5633,7.4227 |
| 135 | 7.0001 | 0.1885 | 1e-04 | 0.1971 | 6.6306, 7.3696 |

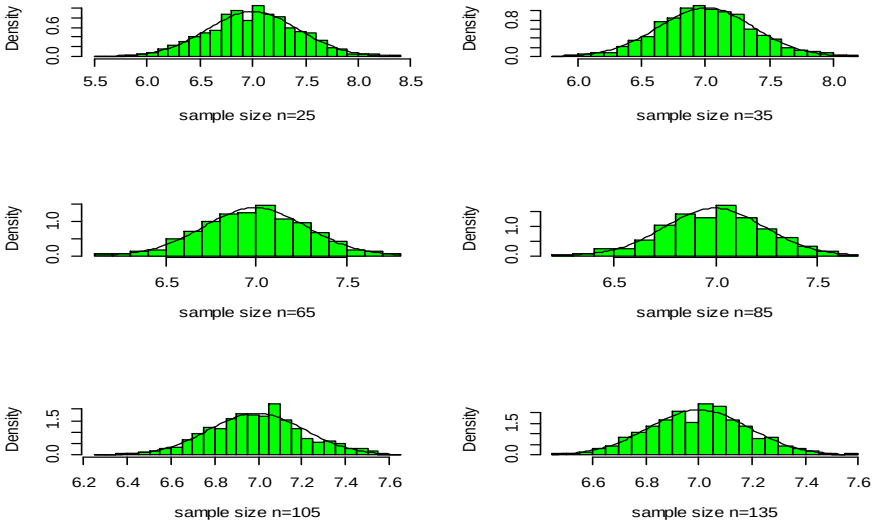


**Figure 2:** Distribution of 1,000 sampling distribution of means when data generated from normal distribution at different sample size $n = (25, 35, 65, 85, 105, 135)$

**Table 3:** The summary statistic of 1,000 sampling distribution of means for Chi-squares distribution $\chi^2$ with 2 df $\mu = 2$ and $\sigma = 2$ at different sample size, $n$

| Sample size $n$ | MSDM =$\mu_{\bar{x}}$ | Standard deviation = $\sigma_{\bar{x}}$ | Bias = $\mu_{\bar{x}}$ - $\mu$ | Formula S.E $\frac{\sigma}{\sqrt{n}}$ | 95% confidence interval |
|---|---|---|---|---|---|
| 25 | 2.0229 | 0.387 | 0.0229 | 0.4 | 1.2644, 2.7813 |
| 35 | 2.0124 | 0.3307 | 0.0124 | 0.3381 | 1.3642, 2.6606 |
| 65 | 2.0072 | 0.249 | 0.0072 | 0.2481 | 1.5192, 2.4952 |
| 85 | 2.0049 | 0.2195 | 0.0049 | 0.2169 | 1.5747, 2.4351 |
| 105 | 2.0041 | 0.1977 | 0.0041 | 0.1952 | 1.6165, 2.3917 |
| 135 | 2.001 | 0.173 | 0.001 | 0.1721 | 1.6619, 2.3401 |

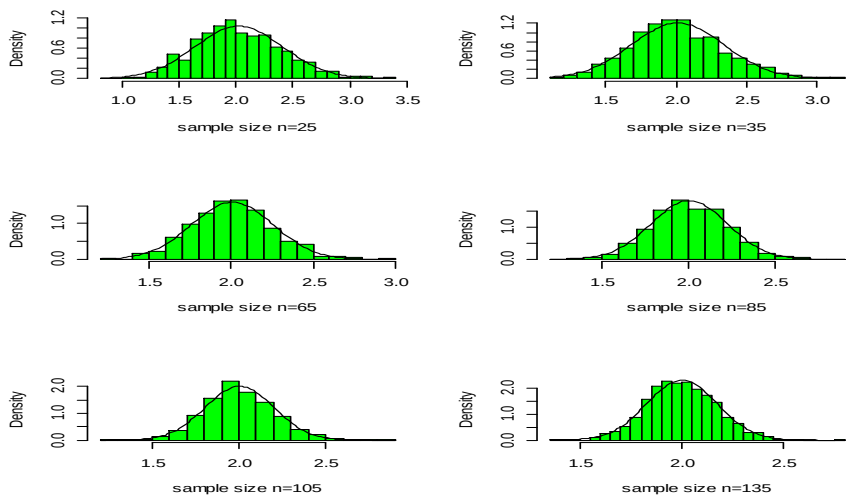


**Figure 3:** Distribution of 1,000 sampling distribution of means when data generated from chi-squares distribution at different sample size $n = (25, 35, 65, 85, 105, 135)$

**Table 4:** The summary statistic of 1,000 sampling distribution of means for Laplace distribution$L(\mu = 4\,, scale\ (b) = 1)$, with population mean $\mu = 4$ and $\sigma = 1.4142$ at different sample size, $n$

| Sample size $n$ | MSDM $=\mu_{\bar{x}}$ | Standard deviation $= \sigma_{\bar{x}}$ | Bias = $\mu_{\bar{x}}$ - $\mu$ | Formula S.E $\frac{\sigma}{\sqrt{n}}$ | 95% confidence interval |
|---|---|---|---|---|---|
| 25 | 3.9949 | 0.4014 | -0.0051 | 0.2828 | 3.2082, 4.7816 |
| 35 | 4.0004 | 0.336 | 4e-04 | 0.239 | 3.3418, 4.6591 |
| 65 | 4.0045 | 0.2536 | 0.0045 | 0.1754 | 3.5074, 4.5016 |
| 85 | 3.9978 | 0.2223 | -0.0022 | 0.1534 | 3.562, 4.4335 |
| 105 | 3.9953 | 0.2009 | -0.0047 | 0.138 | 3.6016, 4.3889 |
| 135 | 3.9931 | 0.1692 | -0.0069 | 0.1217 | 3.6615, 4.3246 |

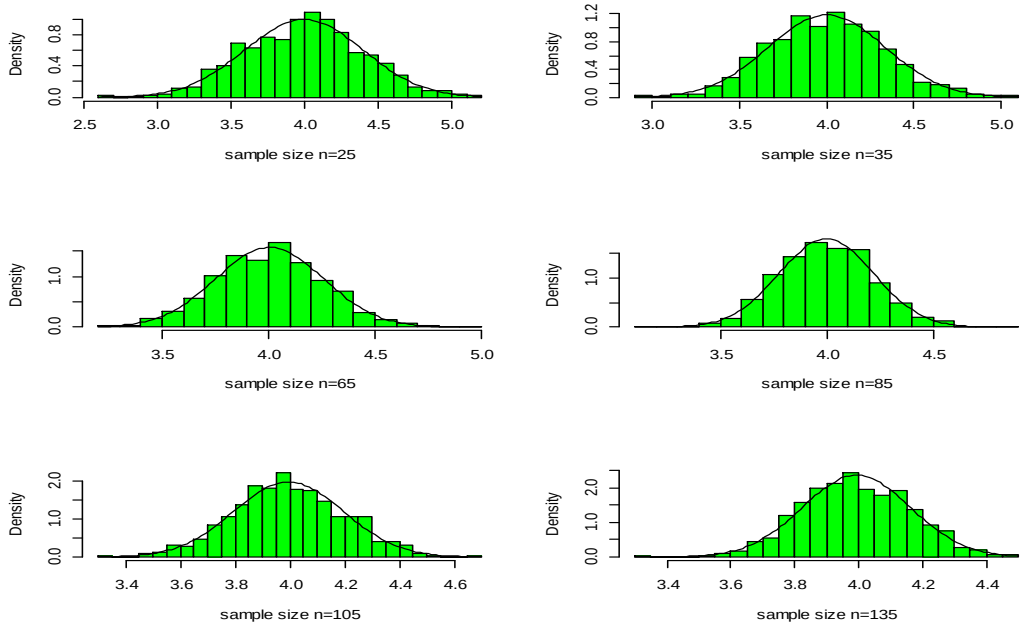


**Figure 4:** Distribution of 1,000 sampling distribution of means when data generated from Laplace distribution at different sample size $n = (25, 35, 65, 85, 105, 135)$

## 4.0 Discussion

In this study, 1,000 repeated random samples of size $n$ = (25, 35, 65, 85,105,135) are generated each time from Normal, Chi-squares, and Laplace distributions, through R code, and the means 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 each distribution such as Normal, Chi-squares, and Laplace 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 true population mean $\mu$ .

For instance, the actual population mean value of the normal distribution $\mu = 7.0$ is attained when sample size $n$ was $n = (25, 35\ and\ 135)$ to 2 decimal places, for Chi-squares, $\mu = 2$ was attained when sample size $n = [\,65, 85, 105\ and\ 135]$ to 2 decimal places, while Laplace with $\mu = 4$ was attained when $n = (\,35, 65)$ to 2 decimal places 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 2, Table 3, and Table 4

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 $\sigma$ 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, as confirmed from Table 2, Table 3, and Table 4. 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 $\mu_{\hat{\theta}}$ is approximately normal ie.$N(\mu, \sigma^2/\mathrm{n})$ this justified use of statistical procedure based on normal distribution theory for construction of interval estimation that contains population mean $\mu$ . As a result of this, the standard normal distribution probabilities table is used in this study to compute 95% confidence interval for population mean μ for each sample size $n$ as a statistical inference, for each distribution considered as confirmed in Table 2, Table 3 and Table 4. Finally, 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 study [1] where the distribution of the sample means look very close to a normal distribution as sample size $n$ increases

## 4.1 Conclusion

In this study, computer simulation has aided visualizing the important abstract concepts namely: sampling distribution of means, normal curve, central limit theorem, standard deviation and standard error 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 population mean,$\mu$. The knowledge and skill obtain through this process is expected to benefit the following such as: students, lecturers, researcher, book writers etc.

## References

**[**1] Ibrahim, S.A., *A Simulation Approach to Studying Normality of Sampling Distribution",* . Al-Hikmah Journal of Pure and Applied Sciences 2016. ( 2): p. 41-121.

[2] Rummerfield, S.A.H.W., *Simulation Methods for Teaching Sampling Distributions: Should Hands-on Activities Precede the Computer?* Journal of Statistics Education, 2020. **28**(1): p. 9-17.

[3] Chandrakantha, L., *Excel Simulation as a Too in Teaching Sampling Distributions in Introductory Statistics* Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS). New York, USA 2014: p. 1-3.

[4] David, P.D., *Using Simulation to Teach Distribution. Journal of Statistics Education.* Journal of Statistics Education, 2004. **12**(1): p. 1-3.

[5] Mills, J.D., *Using Computer Simulation Methods to Teach Statistics: A Review of the Literature.* Journal of Statistics Education 2002. **10**(1): p. 5-12.
[6] Crawley, M.J., *The R Book.* John Wiley & Sons Ltd, England., 2007.
[7] Robert, I.K., *R in action: Data Analysis and Graphic with R. Manning Publications Co., Shelter Island.* 2011.