Research Journal of Mathematical and Statistical Sciences ______________________________ ISSN 2320-6047
Vol. **12(2),** 4-7, September **(2024)** *Res. J. Mathematical and Statistical Sci.*

Short Communication

Central Limit Theorem–An illustration based on simulated data using R

Kalesh M Karun[1*] and Deepthy M.S.[2]
[1]Division of Biostatistics, MOSC Medical College, Ernakulam, 682311, Kerala, India
[2]Department of Biostatistics, Jawaharlal Institute of Postgraduate Medical Education & Research, Puducherry, 605006, India
karunkmk@gmail.com

Available online at: www.iscamaths.com, www.isca.in, www.isca.me
Received 6th April 2024, revised 13th June 2024, accepted 18th July 2024

Abstract

The central limit theorem is the most fundamental theory in modern statistics and quite an important concept in biostatistics, and data science. The central limit theorem states that the sampling distribution of the mean for a variable will approximate a normal distribution regardless of that variable's distribution in the population, when the sample size is large. In real life we cannot repeat studies (resampling) many times to estimate the sampling distribution of the mean. Hence only a simulation-based illustration is possible to understand the concept of central limit theorem. Present study aims to provide a clear understanding of the concept of central limit theorem with the help of simulated data using R codes.

Keywords: Central limit theorem, Distributions, Normal distribution, Simulation, R codes.

Introduction

Distribution of the variable is one of the most important factors in determining the right choice of any statistical analyses. Knowledge regarding the assumption of normality is critical, as it enables the researcher to choose either parametric or non-parametric tests[1]. Prior to statistical analysis, researchers frequently become perplexed regarding the nature of the variables. Many variables like length of hospital stay, time to recovery, haemoglobin level etc., may follow non-normal distributions. Parametric tests are more powerful and yield more precise and accurate estimates than non-parametric tests. So, it is essential to decide whether the data follows normal distribution or not. The normality of the data can be checked either by using Kolmogorov-Smirnov test or Shapiro-Wilk test[2,3]. An alternative option is to plot histograms and see whether the curve fitted is bell shaped symmetrical in nature[4]. When the data fails normality, rather than proceeding straight away with non-parametric tests attempts should be made to make it normal using transformations.

The central limit theorem is one of the most fundamental theories in modern statistics which has contributed greatly to the development of parametric tests. This concept holds significance in biostatistics, mathematics, and data science, making it essential for researchers to understand it[5]. The origin of the central limit theorem states back to Abraham de Moivre's 1738 book, "The Doctrine of Chances"[6]. The mathematical definition of Central Limit Theorem is as follows[7]:

Let $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$ be a sequence of independent random variables, each having the same distribution with finite mean μ and finite variance of σ^2. If $\bar{X}_n$is the mean of $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$, then the distribution of the standardized variable $Z_n = (\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ converges to the normal distribution as $n \rightarrow \infty$.

That means, when sample size is sufficiently large, the sampling distribution of the mean for a variable will approximate a normal distribution regardless of the distribution of that variable in the population[8,9]. In other words, if we randomly take large number of samples with a specific sample size without replacement from a population and plot the means of these samples, the histogram will be bell shaped symmetrical normal curve, when the sample size is sufficiently large.

Usually we conduct a study once, and calculate the mean of that sample. We cannot repeat studies (resembling) many times to estimate the sampling distribution of the mean since it is not feasible and ethical. Hence only a simulation-based illustration is possible to understand the concept of central limit theorem in an easier way. The present article aims to provide the researchers with a clear understanding of the concept of central limit theorem with the help of simulated data using R codes.

Methodology

Probability distributions[10,11] are mathematical functions that model different types of data. To provide a better understanding of the concept of central limit theorem using histogram of sampling distributions we simulated samples from distributions such as Uniform, Exponential, Poisson, Gamma and Binomial.

We generated samples from these five distributions and simulated the data using R software[12] version 4.1.1 and the R code is provided for readers to get a better hands-on experience.

The main steps of data generation and simulation are given below: i. Generate 200 sample units (sample size, n=200) using simulation method, ii. Plot the histogram for the generated sample, iii. Repeat step 1 for various number of replicates such as n=100, n=500 and n=1000, iv. Estimate the statistic (mean) of each of the sample for various replicates, v. Plot the histogram to visualize the shape of the distribution of the statistic (mean) for various situations such as n=100, n=500 and n=1000

Results and Discussion

We created histograms for different distributions including Uniform, Exponential, and Poisson using R code (Figure-1, 3 and 5). It is evident from the plots that none of the generated histograms exhibit a bell-shaped curve, as the samples generated from non-normal distributions. However, it is observed that the histogram plotted for the statistic (sampling distribution) is bell shaped for various selected distributions irrespective of the population/ distribution from which the samples are generated. Also, the shape of the curve became more towards bell shaped symmetrical when the sample size of sampling distribution increases from 200 to 1000 (Figure-2, 4 and 6). The R codes along with observed histograms for various distributions are provided below.

Uniform distribution: Step 1: Generate histogram of uniform distribution based on R code
R code: x=runif(200,1,2);hist(x) # x denote the variable with uniform distribution.
hist(x,prob=T);lines(density(x),col="red",lwd=2)
Step 2: Generate histogram for mean(statistic) obtained from the variable with uniform distribution for different sample sizes such as 100, 500 and 1000. *(The R code given below is only for sample size 1000)*
R code: data=data.frame(replicate(1000,runif(200,1,2))); mean=sapply(data,mean)# number of replicates can change.
hist(mean, prob=T);lines(density(mean),col="red",lwd=2)

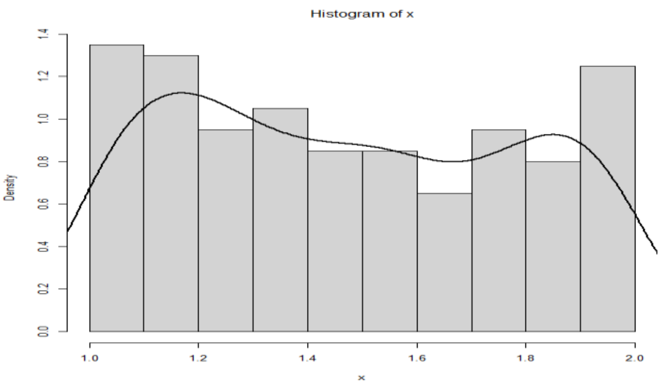

Figure-1: Histogram showing the uniform distribution of variable x.

Exponential distribution: Step 1: Generate histogram of exponential distribution based on R code
R code: x=rexp(200,1/30);hist(x) # x denote the variable with exponential distribution
hist(x,prob=T);lines(density(x),col="red",lwd=2)
Step 2: Generate histogram for mean(statistic) obtained from the variable with exponential distribution for different sample sizes such as 100, 500 and 1000.
R code: data=data.frame (replicate(1000, rexp(200,1/30))); mean=sapply (data,mean)
hist (mean, prob=T);lines(density(mean),col="red",lwd=2)

Poisson distribution: Step 1: Generate histogram of Poisson distribution based on R code
R code: x=rpois(200,8);hist(x) # x denote the variable with exponential distribution
hist(x,prob=T);lines(density(x),col="red",lwd=2)
Step 2: Generate histogram for mean(statistic) obtained from the variable with poisson distribution for different sample sizes such as 100, 500 and 1000.
R code: data=data.frame(replicate(1000,rpois(200,8)));mean= sapply(data,mean)
hist(mean,prob=T);lines(density(mean),col="red",lwd=2)

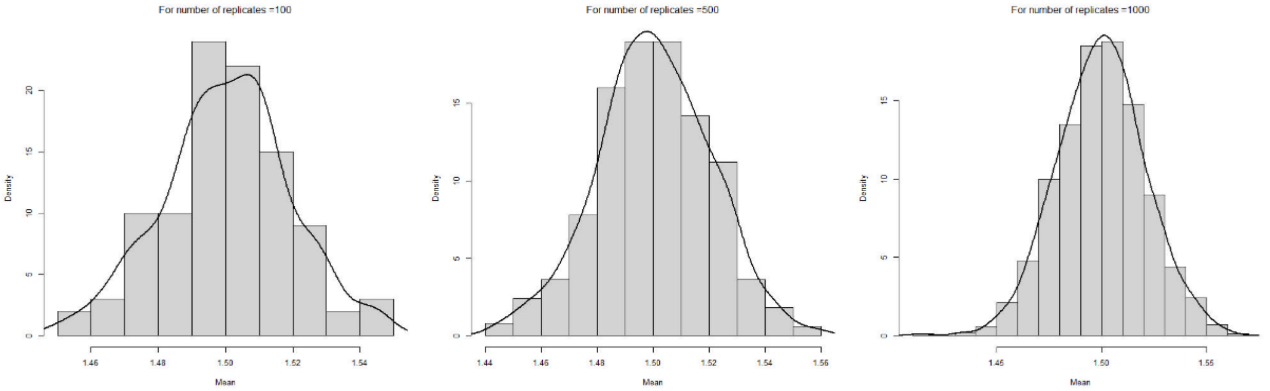

Figure-2: Histogram showing the sampling distribution for number of replicates such as n= 100, n=500 and n=1000 for uniform variable.

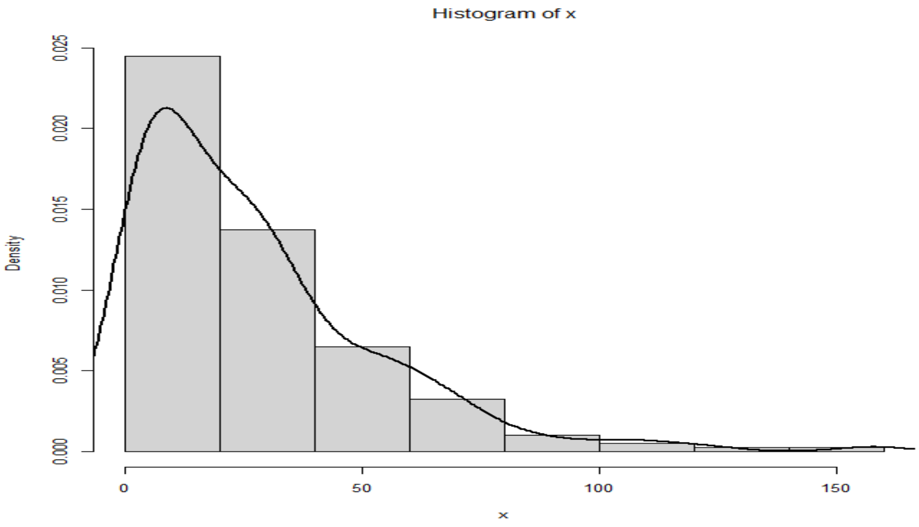

Figure-3: Histogram showing the exponential distribution of variable x.

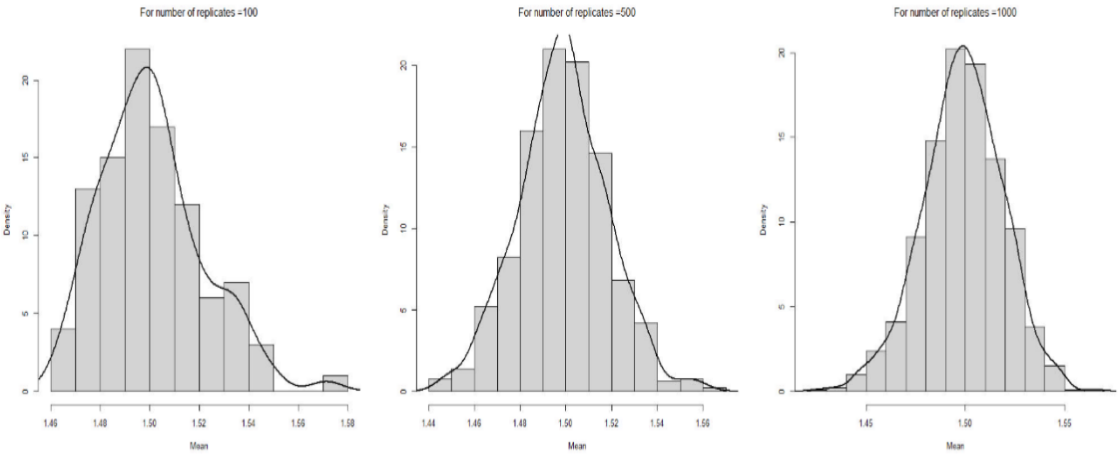

Figure-4: Histogram showing the sampling distribution for number of replicates such as n= 100, n=500 and n=1000 for exponential variable.

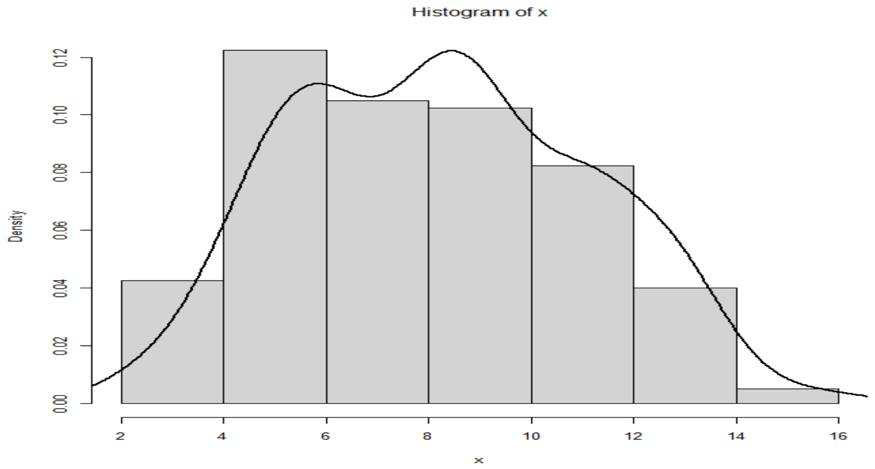

Figure-5: Histogram showing the Poisson distribution of variable x.

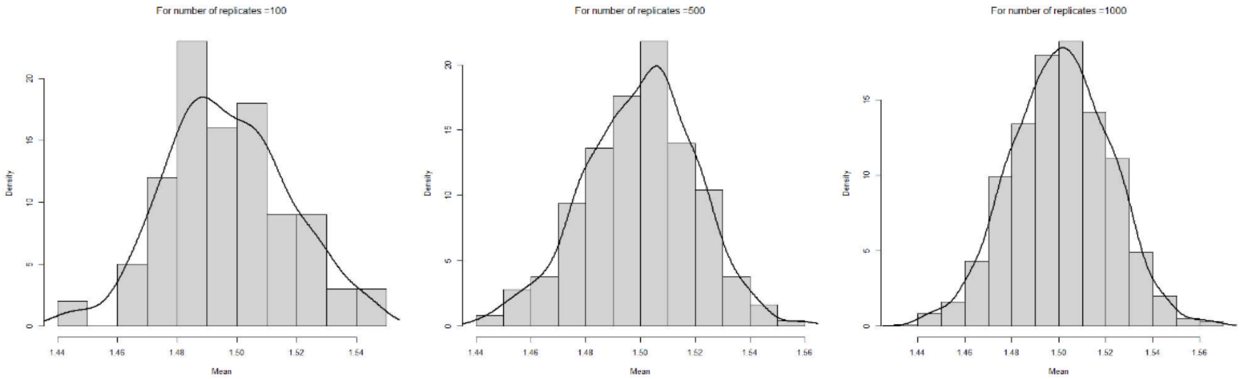

Figure-6: Histogram showing the sampling distribution for number of replicates such as n= 100, n=500 and n=1000 for Poisson variable.

R code for other distributions: R code for other distributions such as Gamma and binomial are given below:

Gamma Distribution: Step 1: Generate histogram of Gamma distribution based on R code
R code: x=rgamma(200,10,4);hist(x)
hist(x,prob=T);lines(density(x),col="red",lwd=2)
Step 2: Generate histogram for mean(statistic)
R code: data=data.frame(replicate(1000,rgamma(200,10,4))); mean=sapply(data,mean)
hist(mean,prob=T);lines(density(mean),col="red",lwd=2)

Binomial distribution: Step 1*:* Generate histogram of binomial distribution based on R code
R code: x=rbinom(200,10,.4);hist(x)
hist(x,prob=T);lines(density(x),col="red",lwd=2)
Step 2: Generate histogram for mean(statistic)
R code: data=data.frame(replicate(1000,rbinom(200,10,.4))); mean=sapply(data,mean)
hist (mean,prob=T);lines(density(mean),col="red",lwd=2)

Conclusion

Central limit theorem plays a crucial role in statistical inference. The histogram plotted for the statistic (sampling distribution) is bell shaped for various selected distributions irrespective of the distribution from which the samples are generated. The shape of the histogram became more towards bell shaped symmetrical when the sample size increases. This simulation-based illustration aids in comprehending the Central Limit Theorem, and this article assists researchers from non-statistical backgrounds in grasping the concept through simulated data facilitated by R code.

References

1. Gerald, B., & Patson, T. F. (2021). Parametric and nonparametric tests: A brief review. *Int J Stat Distrib Appl*, *7*(3), 78-82.
2. Habibzadeh, F. (2024). Data distribution: normal or abnormal?. *Journal of Korean medical science*, 39(3).
3. Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1), 21-33.
4. Rostampour, M., & Azarmi-Atajan, F. (2023). Comparison of normality test methods for some soil properties in the arid land of South Khorasan. *Desert*, 28(2), 381-402.
5. Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology*, 70(2), 144-156.
6. Dunbar, S. R. (2011). The de moivre-laplace central limit theorem. *Topics in Probability Theory and Stochastic Processes*.
7. Glencross, M. J. (1988). A practical approach to the Central Limit Theorem. In *Proceedings of the second international conference on teaching statistics* (pp. 91-95).
8. Adams, W. J. (2009). The life and times of the central limit theorem (Vol. 35). American Mathematical Soc..
9. Rathore, G. S. (2023). Biostatistics And Research Methodology. Academic Guru Publishing House.
10. Pitman, J. (2012). Probability. Springer Science & Business Media.
11. Gerber, S. B., & Finn, K. V. (2005). Basic Ideas of Probability. *Using SPSS For Windows: Data Analysis and Graphics*, 91-94.
12. Moscarelli, M. (2023). Biostatistics With'R': A Guide for Medical Doctors. Springer.