Verification of CLT

MA4240

Savarana Datta Reddy(AI20BTECH11008) Chirag Mehta(AI20BTECH11006)
Dishank Jain(AI20BTECH11011) Vishwanath Sharma(MA20BTECH11010)

IIT Hyderabad

April 25 2022

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Contents

Introduction

Introduction

Central Limit Theorem states that normalised sum of independent and identically distributed random variables tends towards a normal distribution, irrespective of the distribution of random variables.

$$Z = \lim_{n \to \infty} \left(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \right) \tag{1}$$

In this project, we propose to verify the correctness of empirical approximation of Central Limit Theorem by running simulations beginning with a variety of distributions covered in the course.

CLT and empirical approximations

CLT and empirical approximation

While equation (1) suggests that n should be a very large number. In practice, we tend to use the theorem for $n > 30$.
These are some of the assumptions for the distribution:

- The samples drawn are independent.
- The sample size is sufficiently large
- The mean and variance of the sampling distribution are finite

Proof:
Let $X_1, X_2, \ldots, X_n$ be i.i.d random variables with mean μ and variance σ^2. The sum $X_1 + X_2 + X_3 + \cdots + X_n$, has mean $= n\mu$ and variance $n\sigma^2$.
Now consider the random variable

$$Z_n = \frac{X_1 + X_2 + X_3 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}} \tag{2}$$

which is equivalent to

$$Z_n = \sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}} \tag{3}$$

where,

$$Y_i = \frac{X_i - \mu}{\sigma} \tag{4}$$

Each with mean $= 0$ and variance $= 1$.
Since Y_i's are all identically distributed the characteristic equation of Z_n is given as.

$$\phi_{Z_n}(t) = \prod_{i=1}^{n} \phi_{Y_n}\left(\frac{t}{\sqrt{n}}\right) = \left[\phi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n \tag{5}$$

The characteristic equation of Y_1 is given

$$\phi_{Y_1}\left(\frac{t}{\sqrt{n}}\right) = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right), \quad \left(\frac{t}{\sqrt{n}}\right) \to 0 \tag{6}$$

o is the little o notation.
Now the Characteristic equation of z_n in equation (4) is

$$\phi_{Z_n}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \tag{7}$$

We know that

$$\lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = e^x$$

When we apply $\lim_{n\to\infty}$ the equation (6) will change into

$$\lim_{n\to\infty}\phi_{Z_n}(t) = \lim_{n\to\infty}\left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n = e^{\frac{-1}{2}t^2} \tag{8}$$

As all the higher terms will disappear as n goes to higher values. So, The R.H.S will be equal to Characteristic equation of $\mathcal{N}(0,1)$. Therefore as $n \to \infty$ the distribution Z_n will approach $\mathcal{N}(0,1)$. i.e. $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$ converges to the Normal distribution $\mathcal{N}(0,1)$. Hence proved.

Shapiro-Wilk Test

Shapiro-wilk test

- The **Shapiro-Wilk Test** is a test of normality in frequentist statistics. It was published in 1965 by Sameul Sanford Shapiro and Martin Wilk.
- The test statistic(W):

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{9}$$

where

- $x_{(i)}$ is i^{th} order statistic i.e, i^{th} smallest element in the sample.
- $\bar{x}$ is mean of the sample$(x_1, .., x_n)$.
- $(a_1, a_2, .., a_n) = \frac{m^\top V^{-1}}{C}$ where $C = ||V^{-1}m||$.
- Vector m is composed of the expected values of the order statistics of independent and identically distributed random variables drawn from a typical normal distribution.
- V is the covariance matrix of those normal order statistics

P-Value

This test assumes that the population is regularly distributed. If the p value is smaller than the selected alpha level, the null hypothesis is rejected, and evidence that the data being tested are not normally distributed is found. The null hypothesis, on the other hand, cannot be rejected if the p value exceeds the set alpha threshold. Generally we consider alpha threshold as 0.05.

- **If p $\leq$ 0.05**: then the null hypothesis can be rejected (i.e. the variable is NOT normally distributed).
- **If p $>$ 0.05**: then the null hypothesis cannot be rejected (i.e. the variable **MAY BE** normally distributed).

You can find the table for minimum W value required for particular n and p values here.

The Shapiro-Wilk statistic can be calculated using following steps:

- Order the data from the least to greatest. x_j denotes the (j^{th}) order statistic i.e, j^{th} smallest element in the set.
- We calculate value $X_{(n-i+1)} - X_{(i)}$ for i in range [0,$\lceil n/2 \rceil$]
- Get the Shapiro Wilk coefficients(a_{in}) from the table and calculate the value of b.

$$b = \sum_i b_i = \sum_{i=1}^{\lceil n/2 \rceil} (X_{(n-i+1)} - X_{(i)}) a_{in} \tag{10}$$

- Calculate standard deviation(s) of the data set to estimate the Shapiro Wilk statistic (W)

$$W = \frac{b^2}{s^2(n-1)} \tag{11}$$

This estimate of W is close enough to the original one for $n \leq 50$.

Example-1

The following data represents the concentration of nickle in solid waste,

$$58.8, 19, 39, 3.1, 1, 81.5, 151, 942, 262, 331,$$
$$27, 85.6, 56, 14, 21.4, 10, 8.7, 64.4, 578, 637$$

- Arrange in ascending order X=[1, 3.1, 8.7, 10, 14, 19, 21.4, 27, 39, 56, 58.8, 64.4, 81.5, 85.6, 151, 262, 331, 578, 637, 942]
- Difference Matrix $= [941, 633.9, 569.3, 321, 248, 132, 64.2, 54.5, 25.4, 2.8]$
- Shapiro Wilk coefficients for $n = 20$ is [0.4734,0.3211,0.2565,0.2085,0.1686,0.1334,0.1013,0.0711,0.0422,0.0140]
- b value is $= (941)(0.4734) + (633)(0.3211) + .. = 932.8$
- Standard Deviation of the data set $s = 259.72$
- The value of $W = \frac{b^2}{s^2(n-1)} = 0.679$. The minimum value of w needed to have p-value of atleast 0.05 is 0.905 (for n=20). As the observed $W < 0.905$ we can conclude that the given data set is not normally distributed.

Example-2

Consider the following data

$$20, 38, 40, 45, 50, 63, 70, 75, 79, 86$$

- Ascending order $= 20, 38, 40, 45, 50, 63, 70, 75, 79, 86$
- Difference Vector $= 66, 41, 35, 25, 13$
- Shapiro Wilk coefficients for n=10 are $0.5739, 0.3219, 0.2141, 0.1224, 0.0339$
- The value of b $= 62.095$
- The value of $W = 0.959$. The minimum value of w needed to have p-value of atleast 0.05 is 0.869 (for n=10). As the observed $W > 0.869$ we can say that the data set behaves like a normal distribution.

You can find the python code for implementation here.

Hypothesis

Hypothesis

The hypothesis can be framed as follows

H_A : The empirical approximation of the CLT does not hold

H_0 : The empirical approximation of the CLT holds

The significance level α is subject to the method of normality testing. For our case, the Shapiro-Wilk test suggests using a significance level of 0.05.

Procedure

Procedure

We have chosen 4 sampling distributions. For each distribution, we generated 1000 batches of sizes 10, 30, 50 and 100 samples from the sampling distribution. We find the sample mean of each batch and call it $\bar{X}$. From Central Limit Theorem, we know

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{12}$$

To verify the claim, we perform normality tests, which can be classified into two parts

1. Graphical Methods
   - Histogram
2. Frequentist tests
   - Shapiro-wilk test

Distributions used:

Standard Normal: The pdf of standard normal distribution is given by:

$$f_X(x) = \frac{1}{\sqrt{(2\pi)}} exp\left(-\frac{x^2}{2}\right)$$

The mean is 0 and standard deviation is 1. Figure 1 shows the PDF of standard normal distribution.

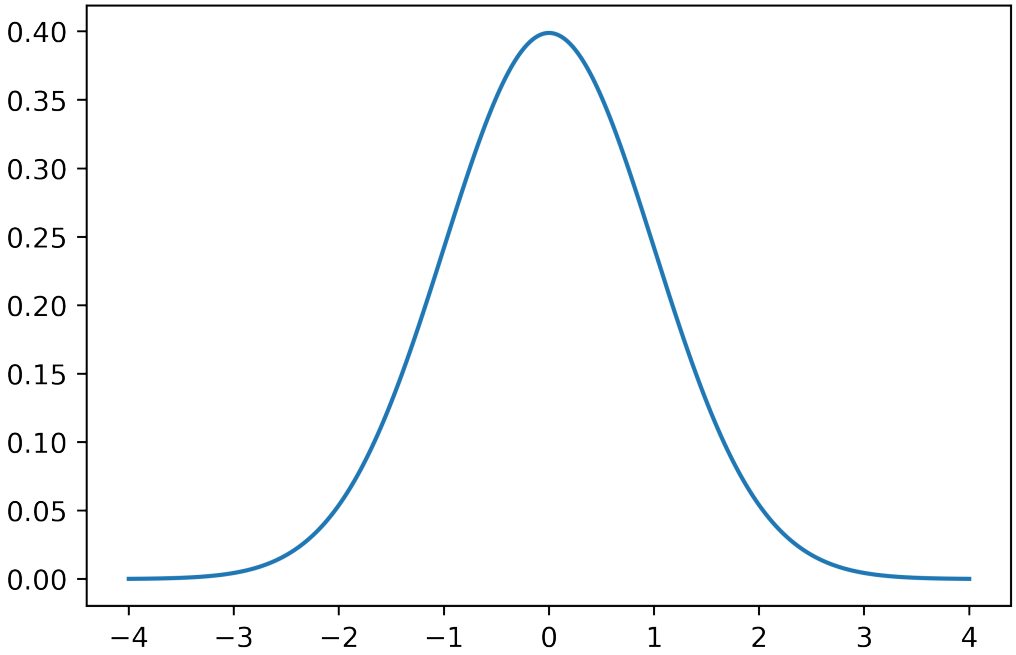

Figure: PDF of standard normal distribution

Continuous uniform distribution: Here, we have used U(0, 1). The PMF is given by

$$f_X(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & otherwise \end{cases}$$

The mean is 0.5 and standard deviation is 0.289. Figure 2 shows the PDF of the uniform distribution.

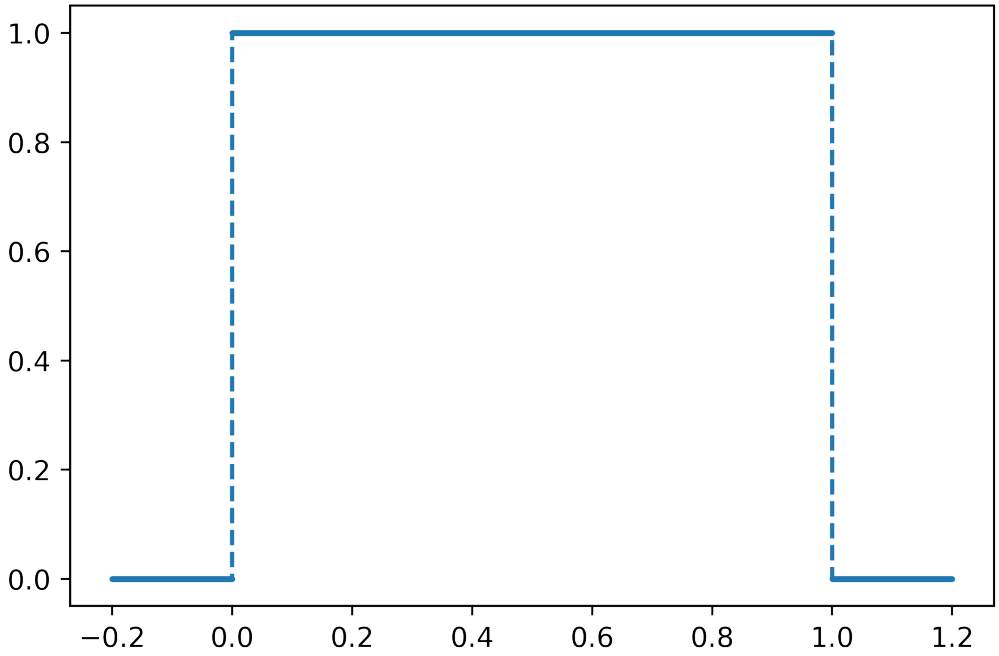

Figure: PDF of uniform distribution

Geometric Distribution: Unlike the other distributions that we used, this distribution is for a discrete random variable. The PMF is given by

$$f_X(k) = (1-p)^{k-1}p,\ k = 1, 2, 3, ...$$

In our experiments, we arbitrarily chose to use $p = 0.35$. The mean is $\frac{1}{p}$ and the standard deviation is $\frac{\sqrt{1-p}}{p}$. The PMF is given in figure 3.

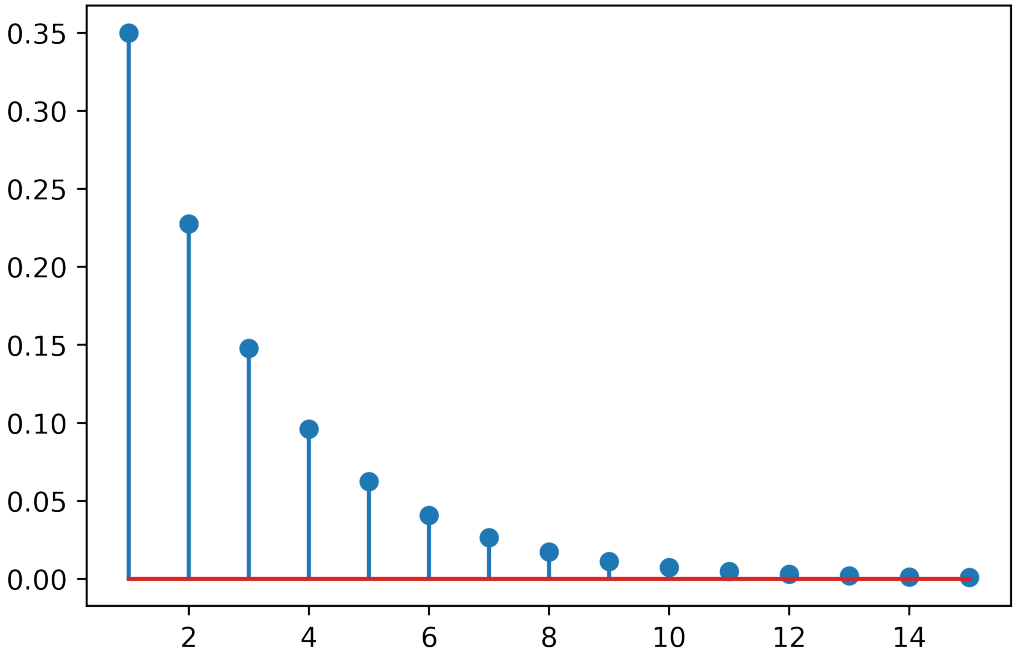

Figure: PMF of geometric distribution

Standard cauchy distribution: The PDF is given by

$$f_X(x) = \frac{1}{\pi(1+x^2)}$$

Neither the mean nor the standard deviation are finite. Thus CLT should not apply on this distribution. Figure 4 shows the PDF of standard cauchy distribution.

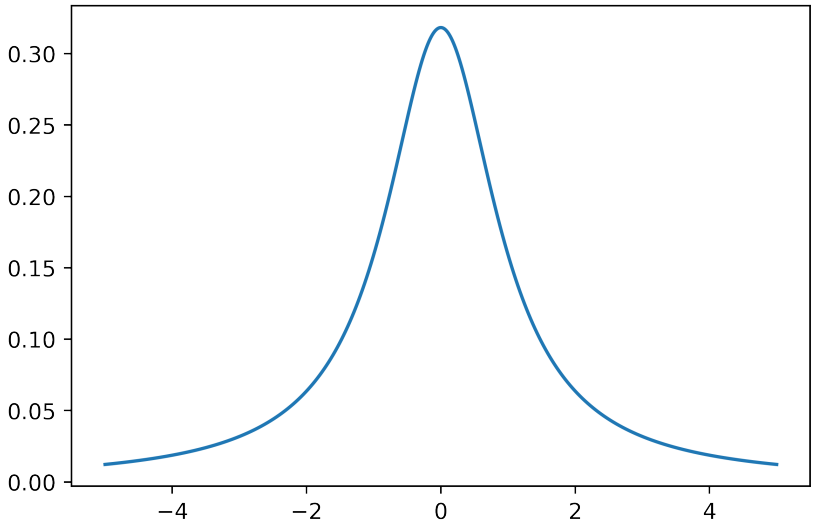

Figure: PDF of Cauchy Distribution

Results

Results

For Normal distribution and uniform distribution, we found that even for the smallest selected sample size of 10, the sample mean follows Normal distribution as per Shapiro-Wilk test. Thus CLT holds true for these for any sample size greater than or equal to 10. For Geometric distribution, we found that CLT did not hold for sample sizes 10 and 30 as per Shapiro-Wilk test. However, CLT held true for sample sizes 50 and 100. For Cauchy distribution, we found that CLT did not hold for any sample size.
You can find the code for simulations here

Conclusion

Conclusion

While using CLT, the empirical approximation is a good one but it may fail. To get better results, one may want to consider using a larger sample size of 50. Also, one may want to verify that the sampling distribution has finite mean and sample variance before applying CLT.

References

References

1. Wikipedia contributors. Central Limit Theorem
2. Wikipedia contributors. Shapiro-Wilk test
3. Wikipedia contributors Cauchy Distribution
4. link-springer. Shapiro-Wilk test (pdf)
5. Shapiro-Wilk test and related tests for normality (pdf)

Thank you!