Verifiation of Central Limit Theorem

Chirag Mehta, G.Savarana Datta Reddy, Dishank Jain, Vishwanath Sharma

Indian Institute Of Technology, Hyderabad

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 Central Limit Theorem by running simulations beginning with a variety of distributions covered in the course.

CLT and empirical approximation

While equation (**??**) 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 $X_1, X_2, \ldots, X_n, \ldots$ 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

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

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

Hypothesis

The hypothesis can be framed as follows

$$H_A : \text{ The empirical approximation of the CLT does not hold}$$
$$H_0 : \text{ The empirical approximation of the CLT holds}$$

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

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

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{9}$$

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 **??** shows the PDF of standard normal distribution.

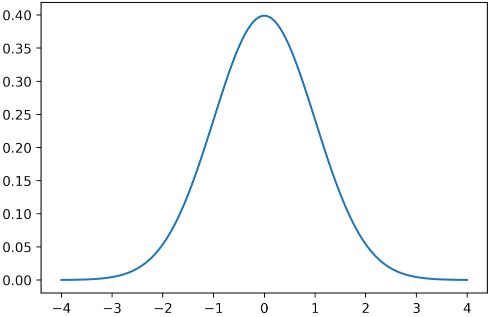

Figure: PDF of standard normal distribution

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

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

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

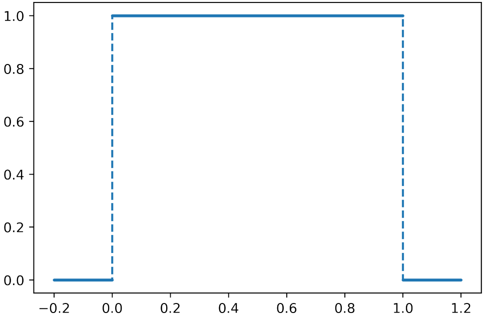

Figure: PDF of uniform distribution

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 $\dfrac{1}{p}$ and the standard deviation is $\dfrac{\sqrt{1-p}}{p}$. The PMF is given in figure **??**.

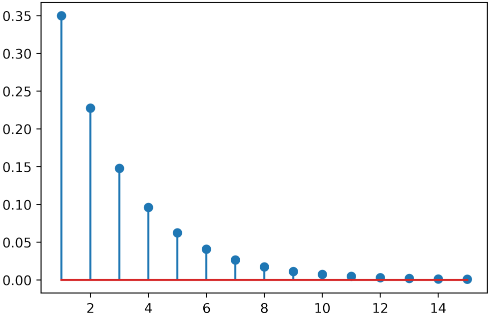

Figure: PMF of geometric distribution

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 **??** shows the PDF of standard cauchy distribution.

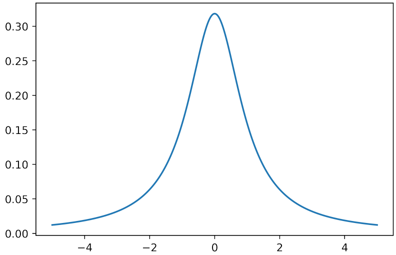

Figure: PDF of Cauchy Distribution

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.

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.

Thank You