Central Limit Theorem

Archi Banerjee
18PH20003

October 2022

1 Introduction

Note: The link to the code is given here: GitHub. Kindly refer to the jupyter notebook to clarify any doubts you might have.

Suppose we have a probability distribution of a random variable X with a given mean μ and standard deviation σ. If we sample n such random variables $X_1, X_2, .., X_n$ from the same probability distribution then the **central limit theorem** states that for $n \lim \infty$ the random variable defined by the sum of the n R.V.s $S = \sum X_i$ follows a **gaussian distribution** with mean μ and standard deviation $\sigma/\sqrt{n}$.

In this assignment, we verify this statement for three underlying base distributions - Gaussian, Exponential and Poisson distributions. For the Gaussian distribution, any number n of random variables will have a mean that follows a Gaussian distribution with the mean and standard deviation prescribed by the Central Limit Theorem. However, for the other two distributions, we have a large cutoff value for n, only beyond which the central limit theorem is satisfied. We also lay out a procedure for attempting to find this cut off value.

2 Gaussian Distribution

For an underlying Gaussian distribution, to demonstrate the robustness of the Central Limit Theorem, we sample $n = 5$, $n = 10$ and $n = 100$ random variables and generate the distribution of their means. We consider a base Gaussian distribution of mean $\mu = 0$ and standard deviation $\sigma = 1$. For $n = 5$, we calculate the value of standard deviation of the distribution of means to be $\sigma = 0.18588$ and the expected value from the central limit theorem is $\sigma = 0.18561$, which are pretty close. For the other two values of n, we also obtain results in agreement with the central limit theorem (for more details, see the .ipynb file attached with this assignment submission).

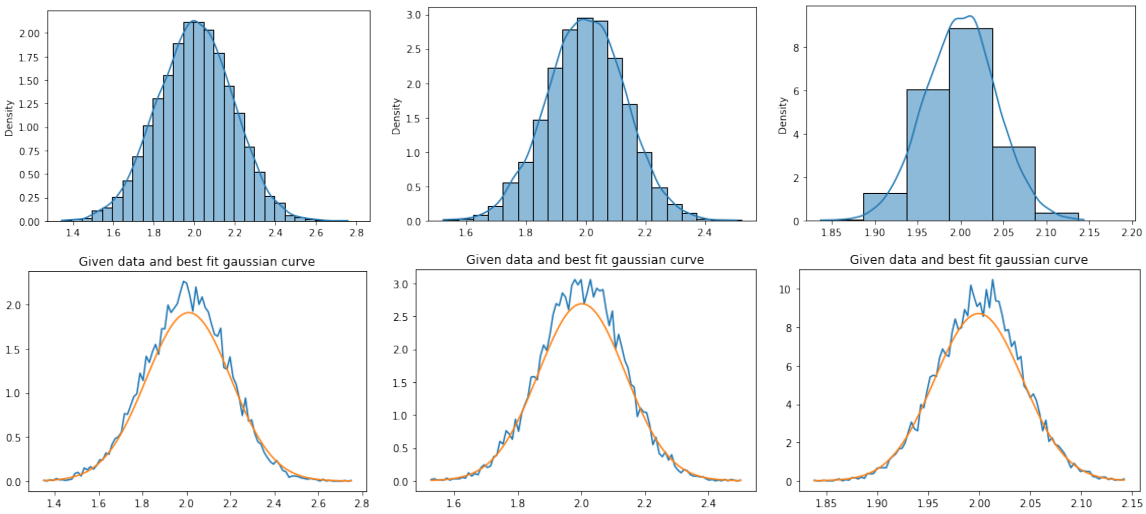
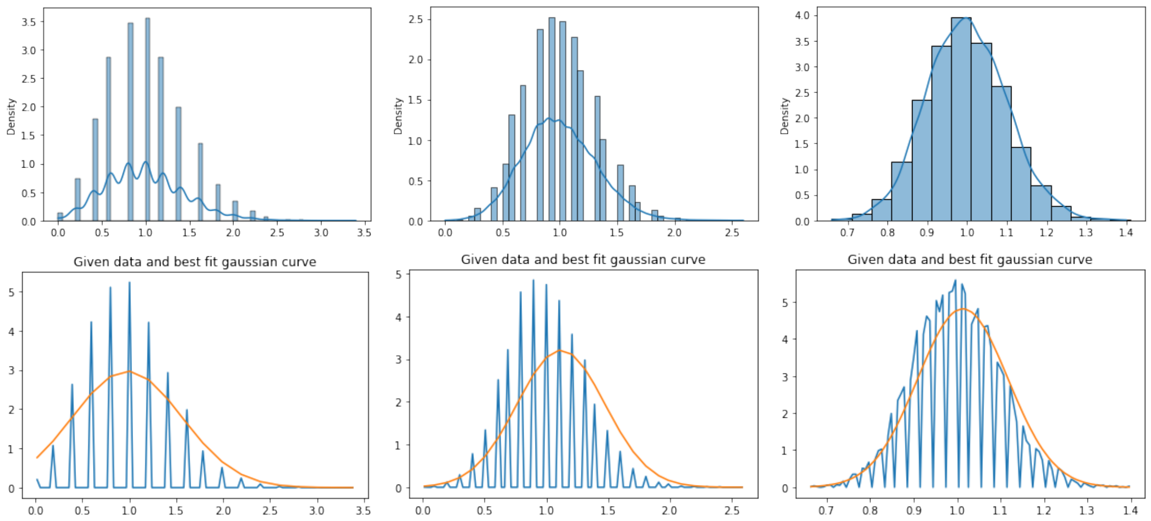


Figure 1: Each column of graphs represents the situation for each value of n. First column is for n = 5, second is for n = 10 and third is for n = 100. As we can see, for both very large (100) and very small values of n (5), the resulting distribution of means is a Gaussian. They are also centered around 0, which was the mean of the original distribution.

3 Poisson distribution

We consider a Poisson Distribution with standard deviation $\sigma = 1$ and mean $\mu = \sigma^2 = 1$. We take $n = 5, 10, 100$ random variables from this distribution and then perform their averaging. We repeat this several times to get a well defined distribution of the mean values. In this case, we get 'spikes' in the distribution because Poisson distributions are discrete - only if we take a large number n of random variables will we get a continuous-looking distribution. As we can see from the figures, the agreement with the Central Limit Theorem is not so good for low values of n. For $n = 100$, the Central Limit Theorem is closer to the



Given data and best fit gaussian curve

Given data and best fit gaussian curve

Given data and best fit gaussian curve

Figure 2: Each column of graphs represents the situation for each value of n. First column is for n = 5, second is for n = 10 and third is for n = 100.

actual distribution. Thus, it is natural to ask if there is a threshold value of n, given σ of the underlying poisson distribution, for and above which the central limit theorem is satisfied. To this end, we have attempted to measure the correctness of the central limit theorem (CLT) by calculating the **mean squared error** between a distribution given by sampling and the distribution it is supposed to mimic according to the CLT (ie, a Gaussian distribution with $\sigma = \sigma_0/\sqrt{n}$). We calculate this mean squared error for underlying poisson distributions of various σ_0 from 0.5 to 2.00 in steps of 0.25. We vary n for each σ from $1 - 176$. We get the following trends of the mean square error for various standard deviations σ_0: As we can see, the

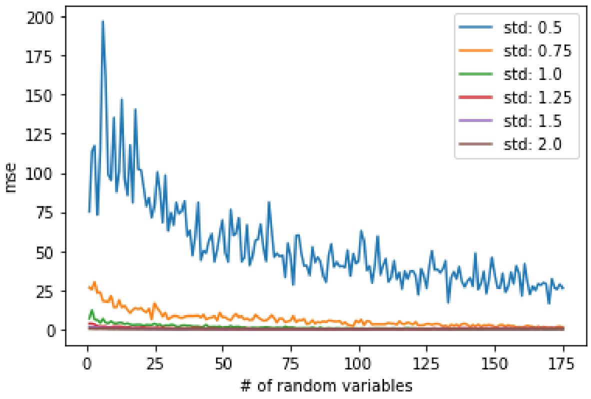


Figure 3: Trend of mean squared error between actual distribution of means and distribution predicted by CLT

mean squared error (between the curves predicted by central limit theorem and the actual distribution of the means) for standard deviation 0.5 (of the base distribution for the random variables) is the highest, and it is the lowest for $\sigma = 2$. It also converges to 0 quicker for higher number of random variables n and for higher σ. One can now easily define a cutoff or a threshold value of n above which the mean squared error is within some infinitesimal range of 0, say $\epsilon \sim 10^{-2}$. We have done this for the case of $\sigma = 1$. The number n for which we obtain such a small mean squared error is $n = 117$, which is roughly of the order of 10^2.

4 Exponential Distribution

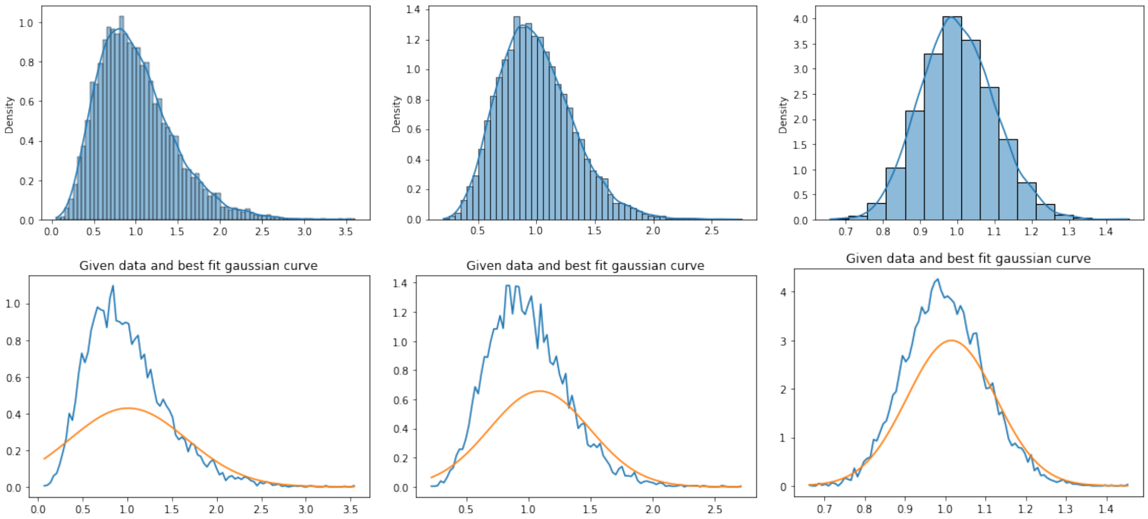


Figure 4: Each column of graphs represents the situation for each value of n. First column is for n = 5, second is for n = 10 and third is for n = 100. The imperfect initial fits are because the distribution is skewed and a symmetric distribution (like a gaussian distribution) cannot fit the data well. This is fixed for large n

As before, the agreement with Central Limit Theorem naturally improves with larger n, as shown in the figure. We employ a similar algorithm as we did for the Poisson distribution to find a threshold value for n beyond which the Central Limit Theorem is practically correct. However, we used the Pearson Correlation Coefficient between the reference gaussian from clt and the condition for convergence was that value of coefficient $r > 0.99$. This gave us a rough estimate of the threshold value ($n \sim 27$ or $10 - 10^2$ order of magnitude). An interesting observation was that the mean squared error was increasing as n was increased, even though the curves were starting to converge. Thus, the correlation coefficient was a better metric.