Central Limit Theorem: Simulation and Applications

St. Xavier's College (Autonomous), Kolkata

- **Name:** Subhajit Karmakar
- **Roll No.:** 410
- **Registration No.:** A01-1112-0837-20
- **Paper Code:** HSTDS6043D
- **Department:** Statistics
- **Session:** 2020 - '23
- **Supervisor:** Prof. Madhura Das Gupta

Declaration: I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

Student's Signature: **Date: 05.04.2023**

Contents

1 Introduction

We know that the behaviour of a phenomenon on a large scale is stable and in most statistical studies, the goal is to identify the characteristics of the population under study. In statistics, the more we draw samples, the more the nature of the population is reflected, necessarily the sample should be a proper representative of the population. Also, in statistics & probability theory, random events exhibit regularity when repeated a large number of times. The **central limit theorem (CLT)** is totally based on this fact only. This central limit theorem is an approximation of the distribution of the sums or means of a large number of random variables, that can be used when the population we are studying is so big that it requires too much money and time to collect data, for example, it is not possible to ask everyone about their age in a community where a large amount of people is residing, to get an idea about the age distribution of that community.

The central limit theorem is based on the concept of convergence in distribution or convergence in law. The theorem is a very key concept in probability theory because the statistical methods and techniques which are applicable to the normal distribution, are also applicable to various types of distributions.

The term **"central"** in the "central limit theorem" is due to the fact that the behaviour of the centre of the distribution of the population gets reflected by the distribution of the sample mean. So, as the sample size increases, the distribution of the sample mean ($\bar{X}_n$) after standardization gets stabilized towards a Normal distribution regardless of the shape of the actual distribution of the population.

The term **"Central Limit Theorem"** was first introduced by **George Poyla** in 1920.

Idea of central limit theorem at a glance through a picture: We will consider the following diagram to understand the central limit theorem visually. We have a population consisting of different units, indicated by various figures in the rectangle. Samples of sizes $S_1, S_2, S_3, \ldots$ are drawn from the population repeatedly and the corresponding sample means are being calculated, denoted by $\mu_1, \mu_2, \mu_3, \ldots$ etc. At the end, we observe the distribution of these sample means and it is clear that the distribution is bell-shaped, symmetric and mesokurtic, which are the properties of the normal distribution. Later we will also verify this by a simulation study.

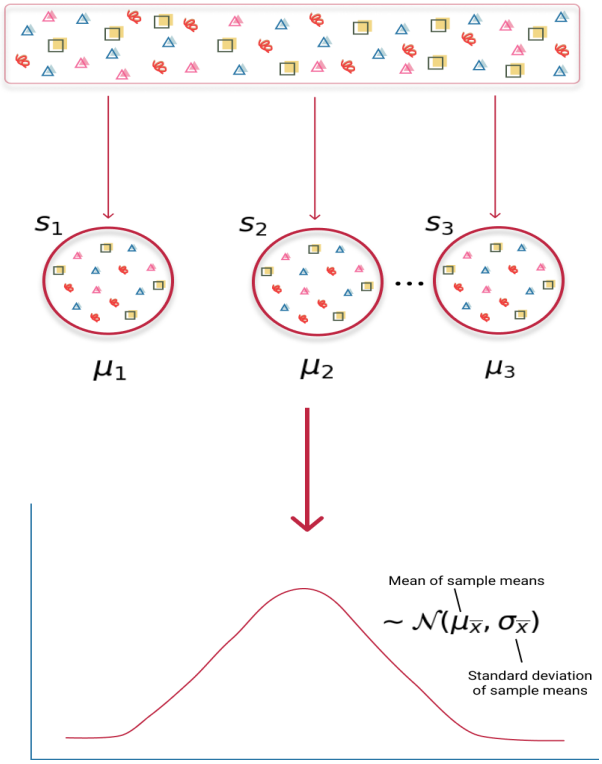

Figure 1

---

2 Brief history of Central Limit Theorem

The Central Limit Theorem (CLT) is one of the fundamental theorems in probability statistics. The CLT has undergone an evolving phase down the centuries. Many mathematicians over the years have proved the CLT in many different cases, therefore providing a different version of the theorem. The origins of the Central Limit Theorem can be traced to the second edition of "The Doctrine of Chances" by Abraham de Moivre published in 1738. Abraham de Moiver's book provided techniques for solving gambling problems and in this book, he provided a statement of the theorem for Bernoulli trials as well as gave a proof for $p = 0.5$. Although de Moivre did not use the term "Bernoulli trials", he wrote about the probability distribution of the number of times "heads" appears when a coin is tossed 3600 times.

3 Definition

Let us formally define the theorem. Let $\{Y_n\}$ be a sequence of random variables. If the distribution of Y_n depends on a parameter 'n' and there exist the quantities a_n and b_n such that,

$$\frac{Y_n - a_n}{b_n} \xrightarrow{L} Z, \qquad as\ n \longrightarrow \infty$$

$$where\ Z \sim N(0,1)$$

Then we say that $\{X_n\}$ obeys CLT.

1. The first statement of the central limit theorem was proposed by **Abraham de Moivre** and **Laplace**, which is as follows. Let us consider a sequence of Bernoulli trials with probability p of success, where $0 < p < 1$. Let S_n denote the number of successes in the first n trials, $n \geq 1$. For any $a, b \in \mathbb{R} \cup \{\pm\infty\}$ with $a < b$. Then,

   $$\lim_{n\to\infty} \left(a \leq \frac{S_n - np}{np(1-p)} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2} dz \quad \text{or,} \quad \frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{L} N(0,1)$$

   This was a very important discovery at that time and it inspired many other mathematicians later to prove it for other cases.

2. A more general and complete proof was given by **Aleksandr Lyapunov**. He proved the central limit theorem under the assumption that $\{X_k\}$'s have finite third-order absolute moments about mean and these moments satisfy some mathematical conditions.
   Let $s_n^2 = \sum_{k=1}^{n} Var(X_k)$ and let $Y = \sum_{k=1}^{n} X_k$. If there exists $l > 0$ such that,
   $\lim_{n\to\infty} \left(\frac{1}{s_n^{2+l}} \sum_{k=1}^{n} E\left[|X_k - E[X_k]|^{2+l}\right]\right) = 0$, then $Z = \frac{Y - E(Y)}{\sqrt{Var(Y)}} \xrightarrow{L} N(0,1)$.

3. The formal definition given by **Lindeberg** is as follows: Let $\{X_n\}$ be a sequence of independent and identically distributed random variables with common mean μ and variance σ^2 $(< \infty)$. Define $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} S_n$. Then,

   $$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{L} Z \sim N(0,1) \qquad or,$$

   $$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{L} Z \sim N(0,1) \quad as\ n \longrightarrow \infty$$

Proof. We will prove the theorem by assuming that the MGF of X_i exists. Let $M(t) = E(e^{tX_i})$ and without loss of generality, we assume that $\mu = 0$ and $\sigma = 1$ (since we end up standardizing $\bar{X}_n$ for the theorem, we might as well standardize the X_i in the first place). Then $M(0) = 1,\ M'(0) = \mu = 1,\ M''(0) = \sigma^2 = 1$

We will just show that the MGF of $\sqrt{n}\bar{X}_n$ converges to the MGF of the $N(0,1)$ distribution, which is $e^{t^2/2}$. By properties of MGFs,

$$E(e^{t(X_1+X_2+...+X_n)/\sqrt{n}}) = E(e^{tX_1/\sqrt{n}})E(e^{tX_2/\sqrt{n}})\dots E(e^{tX_n/\sqrt{n}}) = \Big(M\Big(\frac{t}{\sqrt{n}}\Big)\Big)^n$$

Now, we take logarithm and let $n \longrightarrow \infty$,

$$\begin{aligned}
\lim_{n\to\infty} nlogM\Big(\frac{t}{\sqrt{n}}\Big) &= \lim_{n\to\infty} \frac{logM(yt)}{y^2} && \text{where, } y = 1/\sqrt{n} \\
&= \lim_{n\to\infty} \frac{tM'(yt)}{2yM(yt)} && \text{by L'Hospital rule} \\
&= \frac{t}{2}\lim_{n\to\infty} \frac{M'(yt)}{y} && \text{since, } M(yt) \longrightarrow 1 \\
&= \frac{t^2}{2}\lim_{n\to\infty} M''(yt) && \text{by L'Hospital rule} \\
&= \frac{t^2}{2}
\end{aligned}$$

Therefore, $\lim_{n\to\infty} \Big(M\Big(\frac{t}{\sqrt{n}}\Big)\Big)^n = e^{t^2/2}$.

Thus, $\Big(M\Big(\frac{t}{\sqrt{n}}\Big)\Big)^n$, the MGF of $\sqrt{n}\bar{X}_n$, approaches $e^{t^2/2}$ as desired. Since the MGF determines a distribution uniquely, the CDF of $\sqrt{n}\bar{X}_n$ also approaches the CDF of a standard normal distribution.

Generality of this result: The central limit theorem is a very general result in statistics. The random variable we have defined in the definition can be anything in the real world, so far as long the mean and variance are finite. The distributions of all the real-life data, some of them we can model easily by a suitable distribution, and some are so haphazard that we may not find any suitable distribution to model, but through this theorem, we can always get an idea regarding the true mean of the population when the data itself is very big to handle out.

In the applications of the probability theory, we could have a discrete distribution like the Binomial, a bounded distribution like the Beta, or a distribution with multiple peaks and valleys. No matter what, the act of averaging will cause Normality to emerge. Sometimes, a sample of a small size would accomplish the purpose, sometimes we might need a large amount sample to satisfy the result. For example, if the X_i has a highly skewed or multimodal distribution, we may need 'n' to

be very large before the Normal approximation becomes accurate.

However, if the X_i's are already from a normal population, then whatever the sample size, the distribution of $\bar{X}_n$ will always be $N(\mu, \sigma^2/n)$. No sense of approximation would occur here.

---

4 Simulation study

We should always have a large amount of sample to make the sample mean of a particular distribution having finite mean and variance obey the central law. But while doing this job, a question pops up regarding the sample size, 'How large is sufficiently large ?'. All distributions do not require the same size of the sample, it varies accordingly to the shape and nature of the parent distribution. Though it is not feasible to state properly what sample size is large enough for a good approximation, the only can be said that if the parent distribution is more or less symmetric and relatively short-tailed, then the distribution of sample mean reaches approximately normal distribution for smaller samples than if the parent population is skewed or highly skewed or long-tailed. So, the speed of distribution convergence to the normality also depends upon the nature of the parent population to a large extent.

4.1 Objective

In this simulation study, we will generate random samples of successively higher sizes from different probability distributions with different parameters and compare the convergence towards normality.

4.2 Methodology

First we will draw a random sample of size n from the distribution and compute the sample mean. In this way, we will generate $10,000$ such sample means. Then we will plot a histogram of the standardized sample means. This will be done for various choices of sample sizes n and for different parameter values. From this, we will get an idea about the convergence of the distribution of sample means towards normality. Also, to judge the goodness of fit, we will consider the p-value from the 'Kolmogorov-Smirnov' test (Appendix 11.1). If the p-value is more than 0.05, it means normality is well achieved and vice versa.

4.3 Distributions under study

4.3.1 Exponential(mean $= \theta$)

Let X be a random variable which follows $Exp(mean = \theta)$, then the probability density function of X is given by,

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Here, $E(X) = \theta$ and $Var(X) = \theta^2$. Let $\bar{X}_n$ be the sample mean corresponding to the random sample of size n, where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then, according to the Central Limit Theorem,

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \theta)}{\theta} \overset{a}{\sim} N(0,1) \quad \text{for large n.}$$

Remark: $Exp(mean = \theta)$ is a highly positively skewed distribution.

- **Exp(mean = 0.1):**

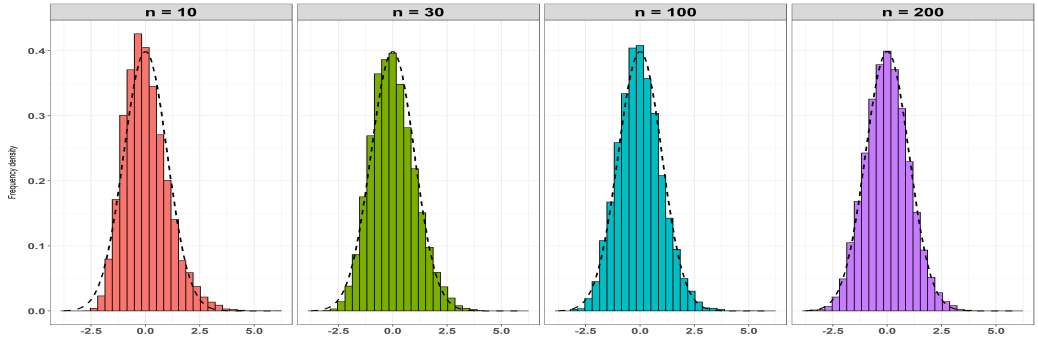

Figure 2

- **Exp(mean = 5):**

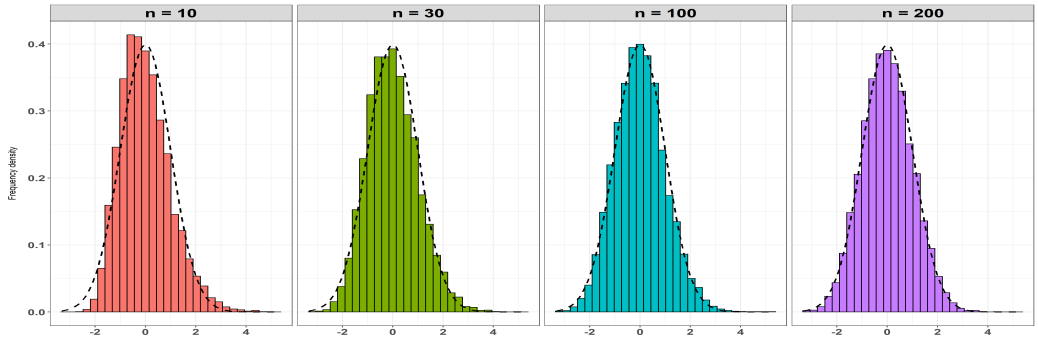

Figure 3

- **Exp(mean = 20):**

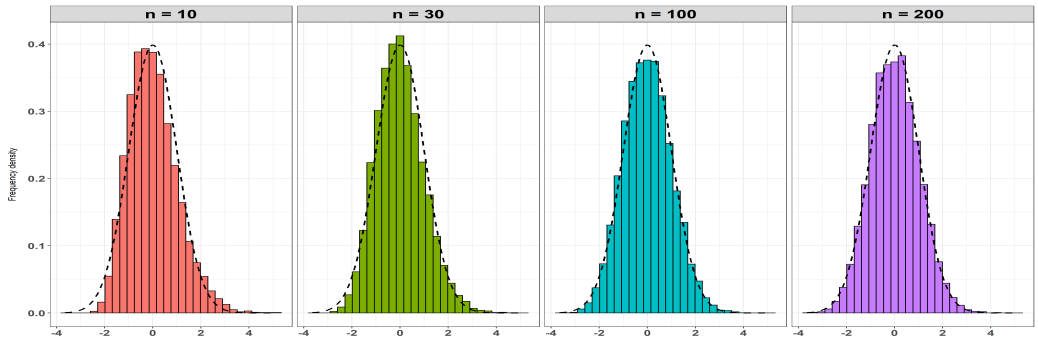

Figure 4

Table of p-value			
n	$\boldsymbol{\theta = 0.1}$	$\boldsymbol{\theta = 5}$	$\boldsymbol{\theta = 20}$
10	0	0	0
30	0.00032	0	0
100	0	= 0.07849	0.0356
200	0.34042	= 0.24759	0.15575

Table 1

Comment: Here, the p-value is decreasing with the increase of parameter value i.e. larger samples are required for converging to a normal distribution.

4.3.2 Beta(a,b)

Let X be a random variable which follows $Beta(a, b)$, then the probability density function of X is given by,

$$f(x) = \begin{cases} \dfrac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Here, $E(X) = \dfrac{a}{a+b}$ and $Var(X) = \dfrac{ab}{(a+b)^2(a+b+1)}$. Let $\bar{X}_n$ be the sample mean corresponding to the random sample of size n, where $\bar{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, then, according to the Central Limit Theorem,

$$Z_n = \frac{\sqrt{n}\left(\bar{X}_n - \dfrac{a}{(a+b)}\right)}{\sqrt{\dfrac{ab}{(a+b)^2(a+b+1)}}} \overset{a}{\sim} N(0,1) \quad \text{for large n.}$$

Remark: $Beta(a, b)$ can be negatively skewed, positively skewed or symmetrical, depending on the parameters.

- **Beta(1,5):**

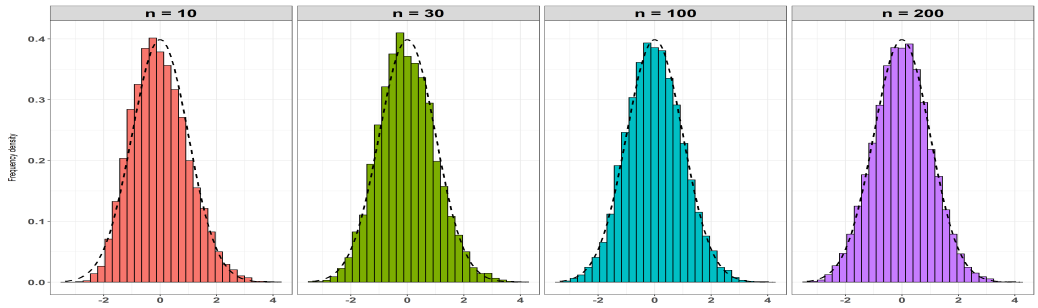

Figure 5

- **Beta(1,10):**

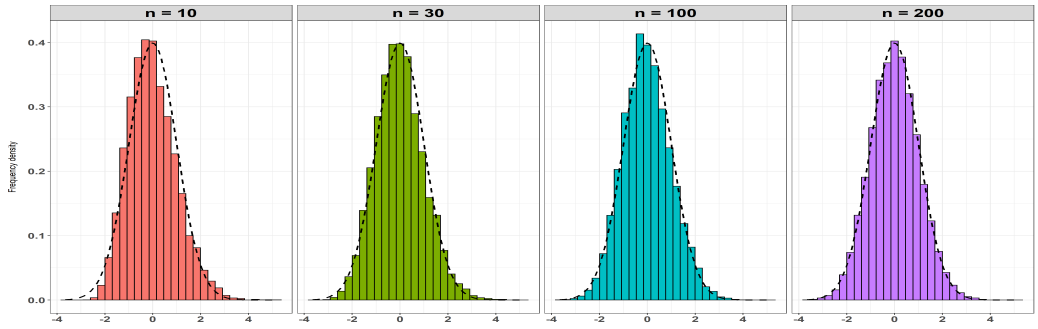

Figure 6

- **Beta(1,20):**

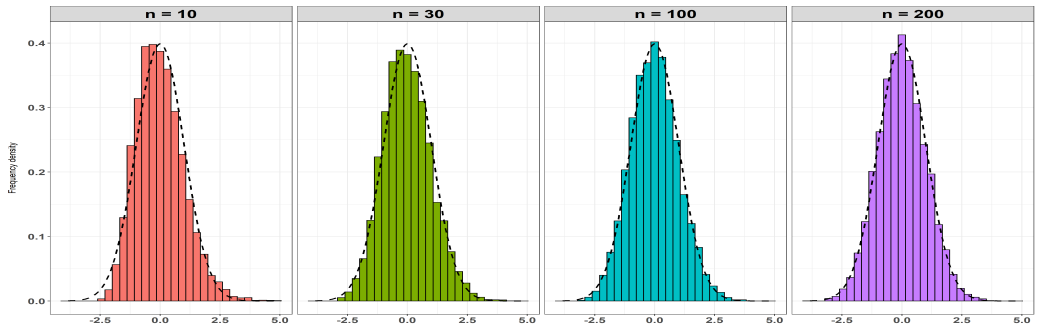

Figure 7

- **Beta(5,5):**

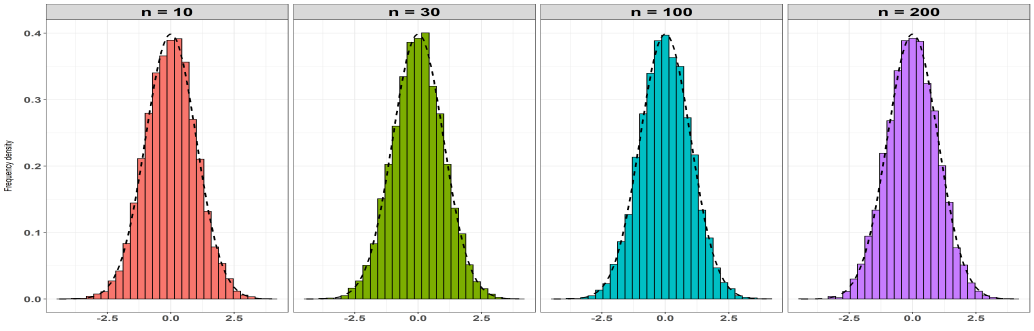

Figure 8

- **Beta(5,10):**

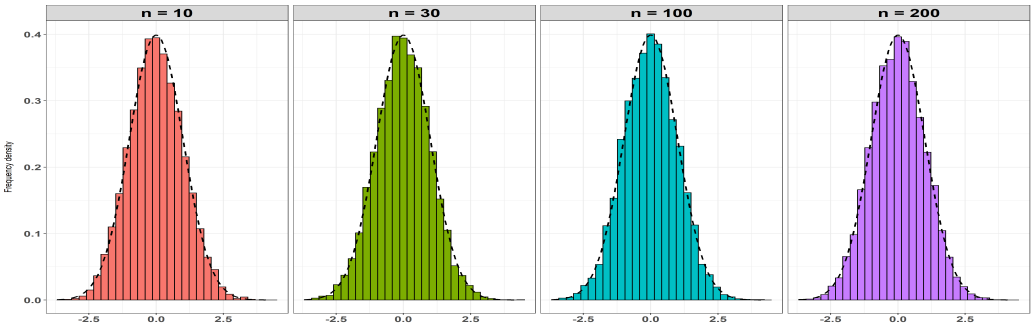

Figure 9

- **Beta(5,20):**

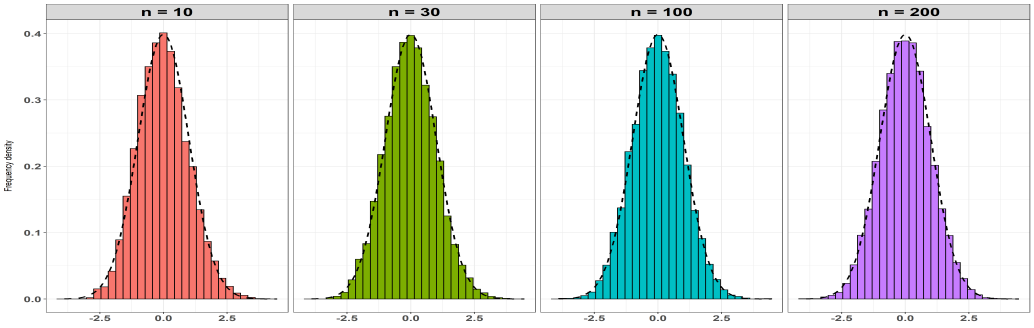

Figure 10

Table of p-value						
n	**a = 1 b = 5**	**a = 1 b = 10**	**a = 1 b = 20**	**a = 5 b = 5**	**a = 5 b = 10**	**a = 5 b = 20**
10	0	0	0	0.33845	0.08294	0.00176
30	0	0.00002	0	0.20336	0.86374	0.10184
100	0.19878	0.00011	0.00417	0.33016	0.19271	0.59972
200	0.99445	0.70942	0.10271	0.71262	0.94459	0.87625

Table 2

Comment: When two parameters are equal, the convergence towards normality occurs more rapidly than the case when the parameters are very different from each other, this is because the parent distribution becomes symmetrical when the two parameters become equal. As the difference between the parameter values is increasing, the normality is occurring for higher sample sizes. Also, it is evident from figures 5 and 9 that if the parameters are shifted by an equal magnitude almost, then the behaviour of convergence does not change. Thus one can get an idea about the nature of convergence towards normality depending on the parameters of the distribution.

4.3.3 Lognormal(μ, σ)

Let X be a random variable which follows $Lognormal(\mu, \sigma)$, then the probability density function of X is given by,

$$f(x) = \begin{cases} \dfrac{1}{\sigma x\sqrt{2\pi}} exp\Big\{ - \dfrac{(ln(x) - \mu)^2}{2\sigma^2} \Big\}, & x > 0 \\ 0, \quad \text{otherwise} \end{cases}$$

Here, $E(X) = e^{\mu+\sigma^2/2}$ and $Var(X) = e^{(2\mu+\sigma^2)}(e^{\sigma^2} - 1)$. Let $\bar{X_n}$ be the sample mean corresponding to the random sample of size n, where $\bar{X_n} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, then, according to the Central Limit Theorem,

$$Z_n = \frac{\sqrt{n}(\bar{X_n} - e^{\mu+\sigma^2/2})}{\sqrt{e^{(2\mu+\sigma^2)}(e^{\sigma^2} - 1)}} \overset{a}{\sim} N(0, 1) \quad \text{for large n.}$$

Remark: $Lognormal(0, 1)$ is a positively skewed distribution.

- **Lognormal(0,0.1):**

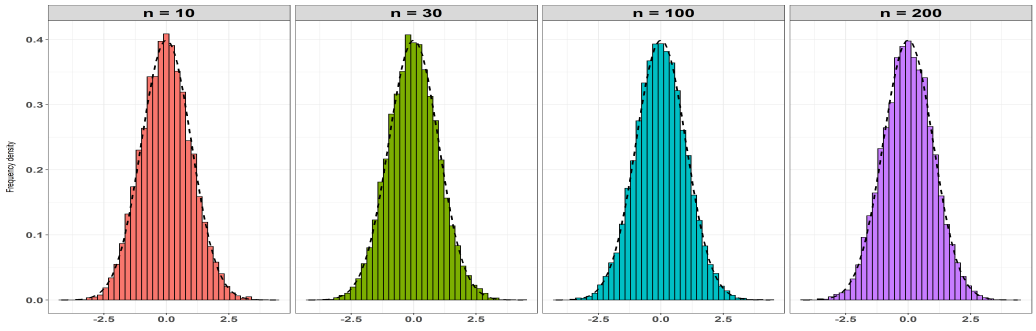

Figure 11

- **Lognormal(0,0.5):**

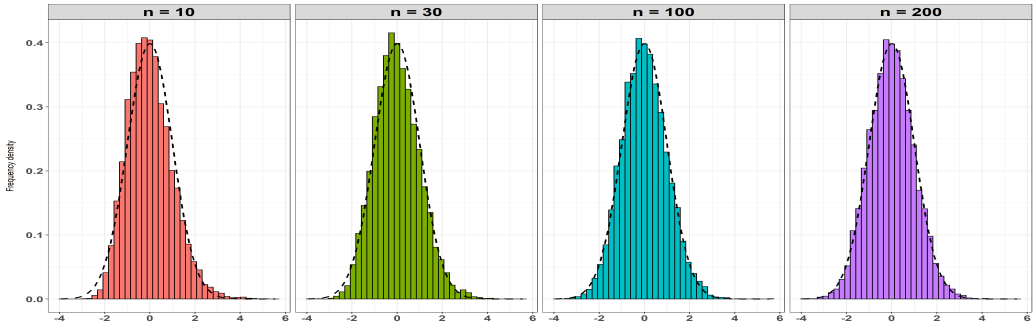

Figure 12

- **Lognormal(0,1):**

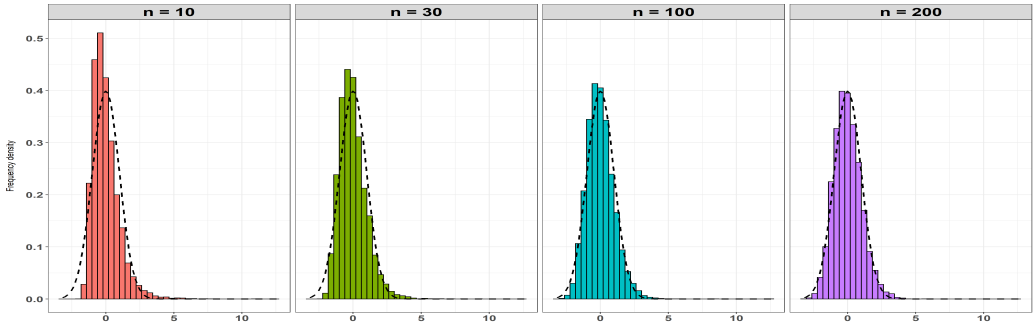

Figure 13

- **Lognormal(5,1):**

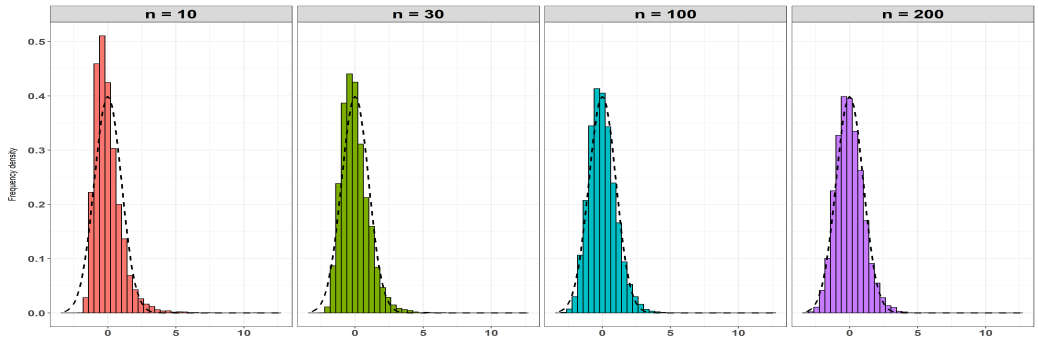

Figure 14

Table of p-value				
n	$\mu = 0 \quad \sigma = 0.1$	$\mu = 0 \quad \sigma = 0.5$	$\mu = 0 \quad \sigma = 1$	$\mu = 5 \quad \sigma = 1$
10	0.33496	0	0	0
30	0.52569	0	0	0
100	0.98492	0.04938	0	0
200	0.86653	0.21163	0	0

Table 3

Comment: As the value of the parameter σ is decreasing, it requires a very larger size of samples for convergence to normality and also from figures 10 and 11, it is evident that the parameter μ does not have any impact on the convergence to normality.

4.3.4 Mixture distribution

Let X be a random variable which follows f_X, where f_X is given by,

$$f(x) = \frac{\alpha}{\sigma_1\sqrt{2\pi}} exp\Big\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\Big\} + \frac{(1-\alpha)}{\sigma_2\sqrt{2\pi}} exp\Big\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\Big\}, -\infty < x < \infty$$

Here,

1. $E(X) = \alpha\mu_1 + (1-\alpha)\mu_2 = \mu$ (say)
2. $V(X) = \alpha(\mu_1^2 + \sigma_1^2) + (1-\alpha)(\mu_2^2 + \sigma_2^2) - \{\alpha\mu_1 + (1-\alpha)\mu_2\}^2 = \sigma^2$ (say)

Where, $\mu_1, \mu_2 \in \mathbb{R}$; $\sigma_1, \sigma_2 \in \mathbb{R}^+$; $0 < \alpha < 1$

Let $\bar{X}_n$ be the sample mean corresponding to the random sample of size n, where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then, according to the Central Limit Theorem,

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \overset{a}{\sim} N(0,1), \quad \text{for large n.}$$

Remark: This distribution has two peak values, in particular, when $\alpha = 0.5$, then it becomes a bimodal distribution.

- $f(\alpha = 0.1, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 5, \sigma_2 = 1)$:

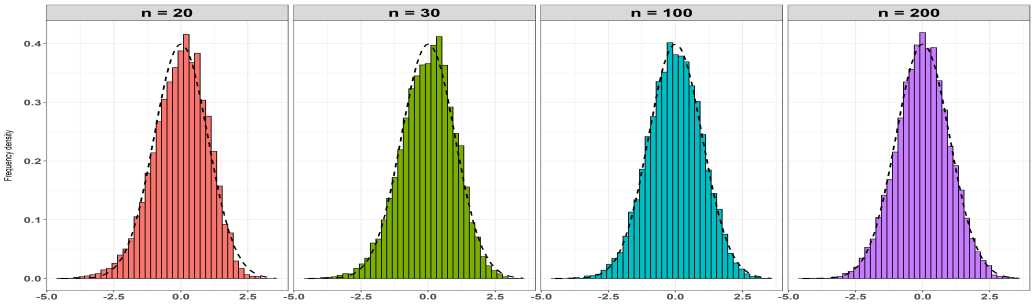

Figure 15

- $f(\alpha = 0.5, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 5, \sigma_2 = 1)$:

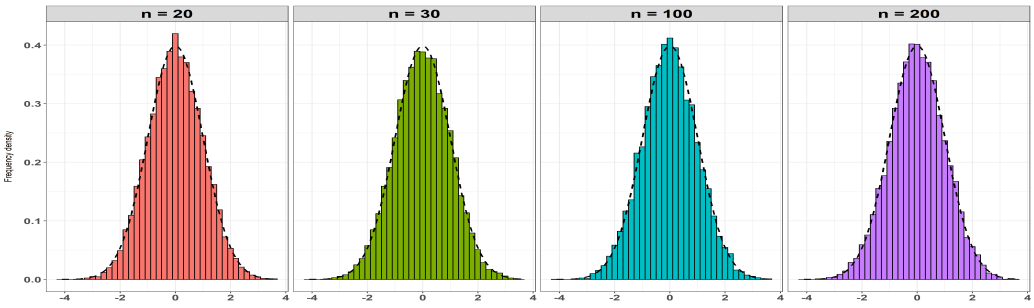

Figure 16

- $f(\alpha = 0.9, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 5, \sigma_2 = 1)$:

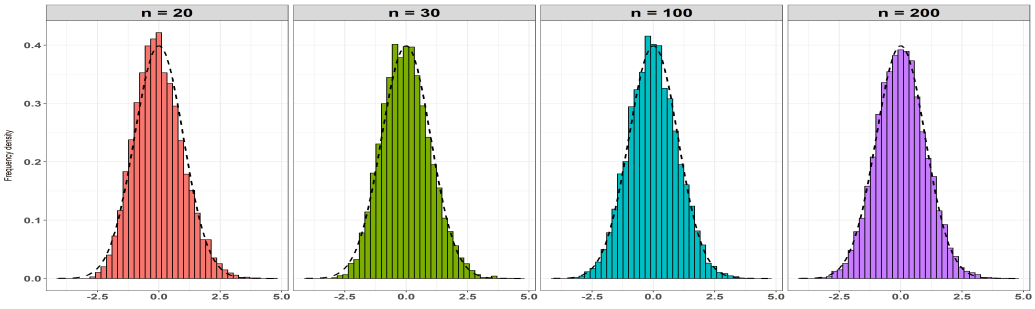

Figure 17

- $f(\alpha = 0.2, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 15, \sigma_2 = 1)$:

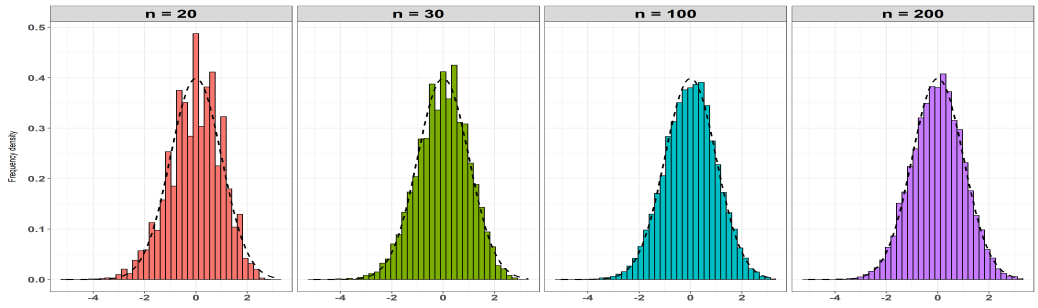

Figure 18

Table of p-value				
n	$\alpha = 0.1$ $\mu_1 = 0$ $\sigma_1 = 1$ $\mu_2 = 5$ $\sigma_2 = 1$	$\alpha = 0.5$ $\mu_1 = 0$ $\sigma_1 = 1$ $\mu_2 = 5$ $\sigma_2 = 1$	$\alpha = 0.9$ $\mu_1 = 0$ $\sigma_1 = 1$ $\mu_2 = 5$ $\sigma_2 = 1$	$\alpha = 0.2$ $\mu_1 = 0$ $\sigma_1 = 1$ $\mu_2 = 15$ $\sigma_2 = 1$
10	0.00001	0.90880	0	0
30	0.00002	0.51402	0.00124	0.00185
100	0.50020	0.19547	0.01122	0.09839
200	0.16010	0.98429	0.20971	0.26634

Table 4

Comment: From the first three columns of the above table, it is evident that as the mixing parameter α departs from 0.5 significantly, it requires samples of relatively larger size to converge to normality.

---

5 A Real-life Application

Let us suppose that we want to study the distribution of the weight of babies in a certain state, say A, at a particular time point, t. Now, obviously, it is not possible for us to survey all the families residing in the state A, since the data is humongous and being a student of statistics, it does not make sense to survey all the babies and calculate the average weight. The central limit theorem is a safeguard in this case, it helps us to balance the time and cost of collecting all the data we need to draw conclusions about the population.

5.1 Methodology

- First, we will draw samples of babies at random from the population. We will draw multiple samples, each consisting of 10 students.
- We will then calculate individual means for these samples. Now, assuming that the distribution of sample means meets Central Limit Theorem's criteria, we can assume the distribution

of the sample means can be approximated to the Normal distribution.

- Then we will plot the histogram of the sample means, the central tendency of this distribution will give us an idea about the true population mean.

Now, to perform this study, let us consider a dataset, each data point denotes the weight (lbs.) of each baby. The picture below shows the first few rows and the dimension of the data.

```
> head(d,10)
       wt
1  6.311835
2  7.251004
3  8.562754
4  8.187968
5  6.375769
6  7.209116
7  8.000575
8  6.812284
9  7.251004
10 6.250105
> dim(d)
[1] 106251      1
>
```

5.2 Distribution of the data

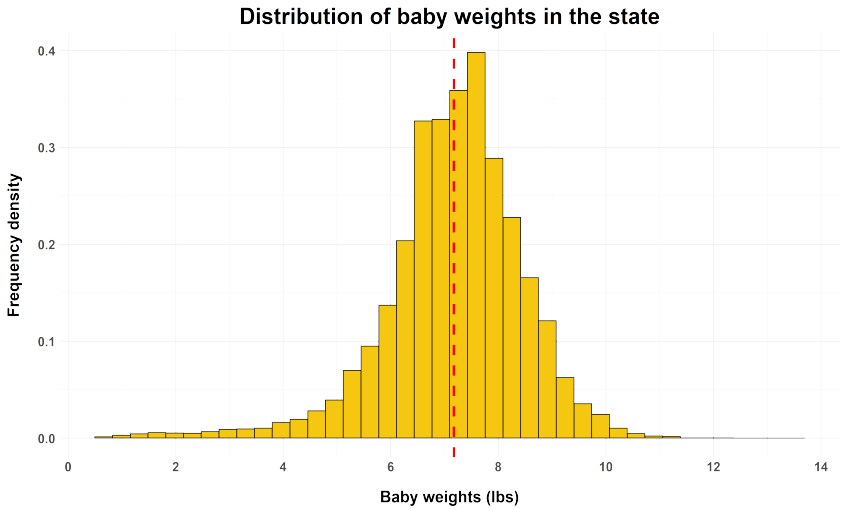

Figure 19

The red line indicated in figure 19 is the population mean, i.e. the true mean weight of the babies is 7.171568 lbs. From the plot, it is clear that the distribution is not really symmetrical, and has a long left tail. So the population distribution does not possess the characteristic of normal distribution at all.

5.3 Distribution of sample mean

Next, we observe the distribution of sample means by considering the sample of size 10, however, it is a very small sample size and practically we should draw samples of size nearly 1000 (10 per cent of the total population).

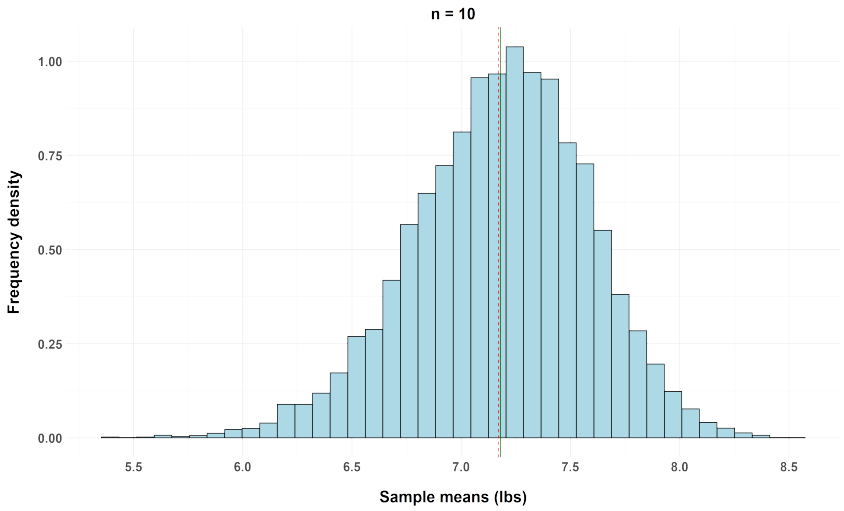

Figure 20

It is reflected in figure 20 that the distribution is well symmetric for sample size 10 only but the dispersion is too high which indicates we might commit a large amount of error while prescribing the estimate of the mean. The mean of this sample mean is coming out as 7.180109 which is very close to the population mean.

Now we perform the same study by considering samples of successively higher sizes.

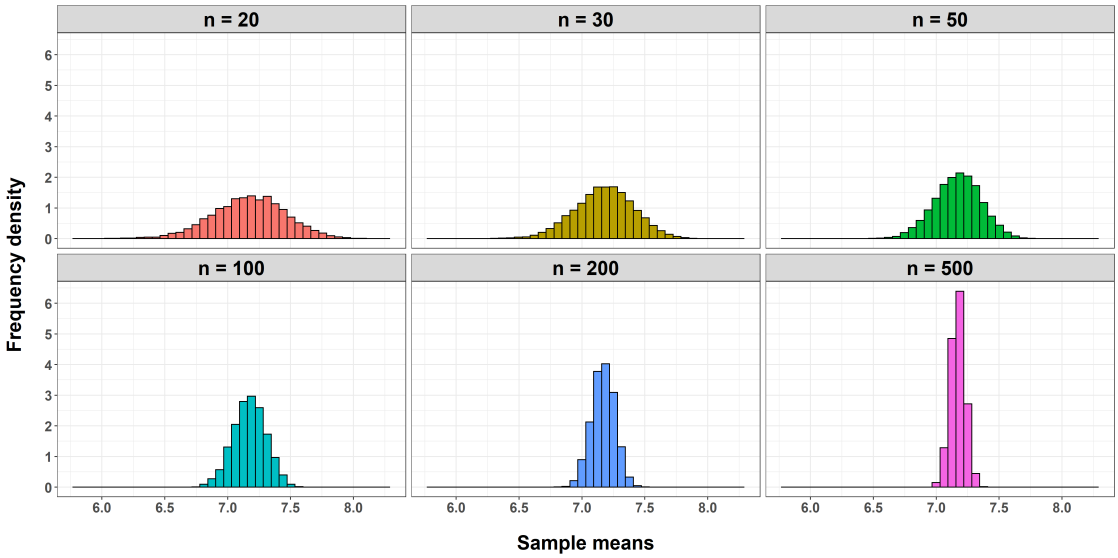

Figure 21

Now, from figure 21, we observe that as the sample size increases, the closer we get to the shape of a Normal distribution, the higher number of samples also reduces the variability in the sampling distribution. Looking back to the definition of the Central Limit Theorem, the standard deviation of the sampling distribution, also called standard error, is equal to $\sigma/\sqrt{n}$ and so the standard error

has decreased substantially for $n = 500$. Thus we will be having a very small amount of error while giving an estimate of the population mean.

5.4 Conclusion

Sample size	means
20	7.169231
30	7.175709
50	7.170494
100	7.172156
200	7.168829
500	7.171646

The means of the sample means for various choices of n are reported in the picture. The population mean is 7.17156 lbs and thus we can justify how close are the sample means to the population mean. We have obtained an approximate of population mean which is nearly identical to the true mean, based on 200 or 500 samples only and even not knowing any details about the original distribution which consists of 106251 units (very large) in total. This is the power of Central Limit Theroem.

---

6 Hypothesis testing

Here we will check the behaviour of the power of a hypothesis test if we use the asymptotic distribution of sample quantile instead of the exact distribution of the test statistic. The level of significance (α) is 0.05 in both cases.

Here we will test for the location parameter of a **Cauchy(μ,1)** distribution. So, the testing problem can be written as below:

$$\text{To test,} \quad H_0 : \mu = 0 \quad \text{against} \quad H_1 : \mu > 0$$

6.1 Methodology

Here we will consider a random sample of size n where, $n = 161$, which is generated from $Cauchy(0.2, 1)$. Purposefully, we have considered a sample of odd size to avoid computational problem in obtaining the distribution of the test statistic (sample median). Now, we will perform the test by considering both exact and asymptotic distribution of the test statistic and will draw the power curves. In both the approaches, the sample observations remain the same for comparison purpose. Also, in this testing problem, We will use the sample median($\widetilde{X}$) as our test statistic since it is consistent, unbiased and also the MLE of μ, so it is a good test statistic.

6.2 Approach 1: Exact distribution

Here, we will reject the null hypothesis at α level of significance if $\widetilde{X} > c$, where c is such that, $P(\widetilde{X} > c) = \alpha$, under H_0. Our next task is to detect the value of c. Thus, the probability density function of $\widetilde{X}$ or the $(k+1)^{th}$ order statistic is given by:

$$f_{X_{(k+1)}} = \frac{(2k+1)!}{k!k!}\{F(x)\}^k f(x)\{1-F(x)\}^k, \quad x \in \mathbb{R} \quad \cdots\cdots(1)$$

$$\Big[\text{Where, } F(\cdot) : \text{CDF of } C(\mu, 1); \quad f(\cdot) : \text{PDF of } C(\mu, 1)\Big]$$

In our case, 2k+1 = 161 $\Rightarrow$ $\boxed{\text{k = 80}}$

Thus, c is the solution of the equation,

$$\int_c^{\infty} f_{X_{(k+1)}}(x)dx = \alpha$$

$$\Rightarrow \frac{(2k+1)!}{k!k!} \int_c^{\infty} \{F(x)\}^k f(x) \{1 - F(x)\}^k dx - \alpha = 0 \quad \cdots\cdots (2)$$

We have, under H_0, $f(x) = \dfrac{1}{\pi(1+x^2)}$ and $F(x) = \dfrac{1}{2} + \dfrac{1}{\pi} tan^{-1}(x)$

Clearly, the above equation consists of only one variable c, so solving the above non-linear equation numerically, we get the value of c as, **0.205274**. So, the critical region is given by:

$$w : \{X | \, \widetilde{x} > 0.205274\}$$

Where X is the vector of any random sample and $\widetilde{x}$ is the corresponding sample median.

Therefore, the power function for the above test is given by:

$$\beta_{\mu}^{(1)} = P(\widetilde{X} > 0.205274 \mid \mu) = \int_{0.205274}^{\infty} H(x|\mu)dx \quad \cdots\cdots (3)$$

$$\Big[\text{Where, } H(x) = f_{X_{(k+1)}} = \frac{(2k+1)!}{k!k!} \{F(x)\}^k f(x) \{1 - F(x)\}^k,$$
$$\text{and, } f(x) = \frac{1}{\pi\{1 + (x-\mu)^2\}}; \quad F(x) = \frac{1}{2} + \frac{1}{\pi} tan^{-1}(x - \mu)\Big]$$

Now, for different choices of $\mu(> 0)$, we will get the power values of the test from the expression (3).

6.3 Approach 2: Asymptotic distribution (Using CLT)

Here we will use the asymptotic distribution of the sample median to perform the test. When the sample size (n) is very large, by using the asymptotic distribution of sample quantile (derived by using CLT, Appendix 11.2), the asymptotic distribution of sample median $(\widetilde{X})$ is given by,

$$\sqrt{n}\Big(X_{(n/2)} - \xi_{1/2}\Big) \overset{a}{\sim} N\Big(0, \frac{1/2(1-1/2)}{\{f_X(\xi_{1/2})\}^2}\Big)$$

$$\Rightarrow \sqrt{n}\Big(X_{(n/2)} - \xi_{1/2}\Big) \overset{a}{\sim} N\Big(0, \frac{1}{4\{f_X(\xi_{1/2})\}^2}\Big) \quad \cdots\cdots (4)$$

Now, if $X \sim C(\mu, 1)$, we have,

$$f_X(\xi_{1/2}) = f_X(\mu) = \frac{1}{\pi\{1 + (\mu - \mu)^2\}} = \frac{1}{\pi}$$

Therefore, from (5), we get,

$$\sqrt{n}\Big(X_{(n/2)} - \xi_{1/2}\Big) \overset{a}{\sim} N\Big(0, \frac{\pi^2}{4}\Big)$$

$$\Rightarrow X_{(n/2)} \overset{a}{\sim} N\Big(\mu, \frac{\pi^2}{4n}\Big) \quad \cdots\cdots (5)$$

Here also, we reject our null hypothesis at α level of significance if $X_{(n/2)} > c$, where, c is such that, $P(X_{(n/2)} > c) = \alpha$, under H_0. We have to determine the value of 'c'.

From (5), we have,

$$P(X_{(n/2)} > c) = \alpha$$

$$\Rightarrow P\Big(\frac{X_{(n/2)} - \mu}{\sqrt{\pi^2/4n}} > \frac{c - \mu}{\sqrt{\pi^2/4n}}\Big) = \alpha$$

$$\Rightarrow 1 - \Phi\Big(\frac{\sqrt{4n}c}{\pi}\Big) = \alpha \quad \Big[\text{Since, } \mu = 0, \text{ under } H_0\Big]$$

$$\Rightarrow \frac{\sqrt{4n}c}{\pi} = \Phi^{-1}(1 - \alpha) = \Phi^{-1}(0.95) = 1.644854$$

$$\Rightarrow c = 1.644854 \cdot \frac{\pi}{\sqrt{4n}} = 1.644854 \cdot \frac{\pi}{\sqrt{644}} = 0.2036265$$

Thus, the value of c we have got is **0.2036265**.

Therefore, the **power function** is given by:

$$\begin{aligned}
\beta_\mu^{(2)} &= P(X_{(n/2)} > 0.2036265 \mid \mu) \\
&= 1 - P(X_{(n/2)} < 0.2036265 \mid \mu) \\
&= 1 - P\Big(\frac{X_{(n/2)} - \mu}{\sqrt{\pi^2/4n}} < \frac{c - \mu}{\sqrt{\pi^2/4n}}\Big) \\
&= 1 - \Phi\Big(\frac{\sqrt{4n}(c - \mu)}{\pi}\Big) \\
&= 1 - \Phi(8.077799(0.2036265 - \mu)) \quad [\because n = 161] \\
&= 1 - \Phi(1.644854 - 8.077799\mu) \\
&= \Phi(8.077799\mu - 1.644854) \quad \cdots\cdots (6)
\end{aligned}$$

Now, for different choices of $\mu(> 0)$, we will get the power values of the test from the expression (6).

Now, we plot the power curves obtained by two different methods, which are shown below:

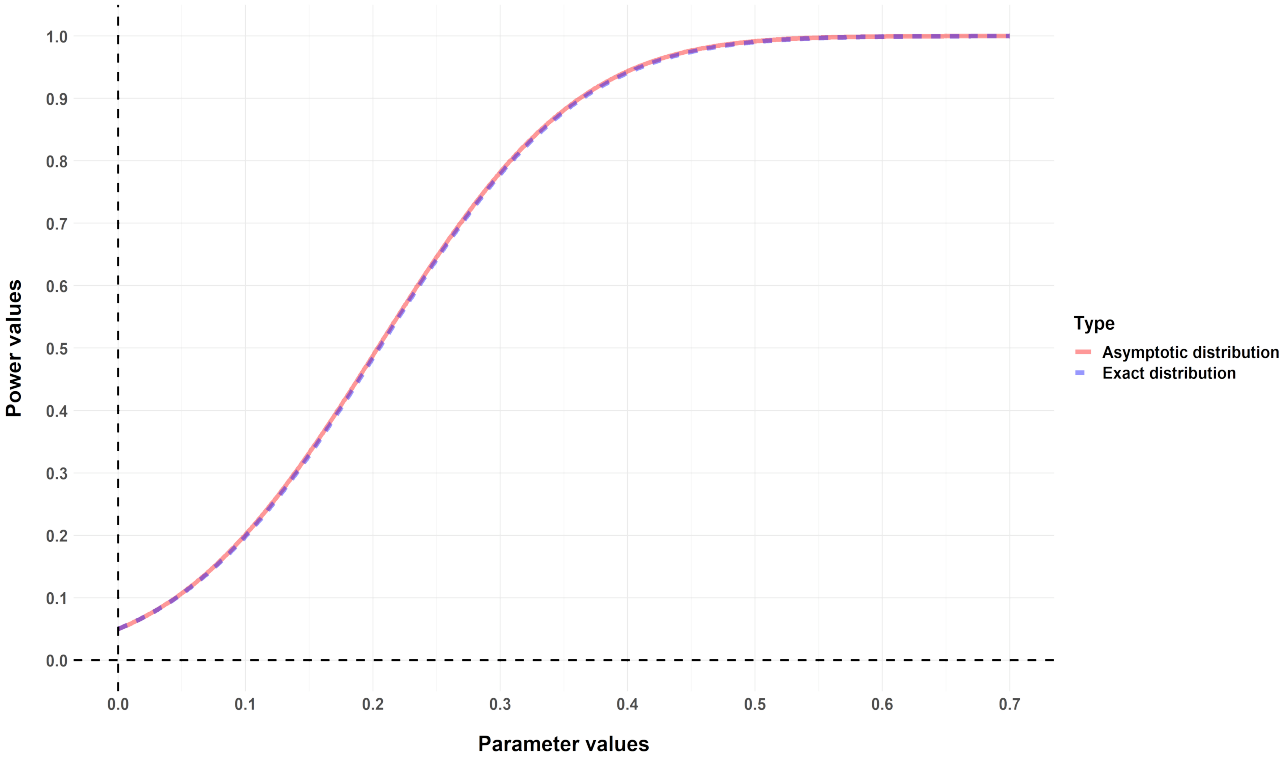

Figure 22

6.4 Conclusion

From the above plot, we see that the power curves obtained by the two approaches are almost identical. Now it is evident that the expression of the power curve obtained through the first approach is very difficult to understand and also is very problematic to work with but if we go by the second approach, all we require is a sample of a quite larger size and then Central Limit Theorem handles everything and both of them yield almost identical results. In this aspect, CLT helps us a lot in our work.

---

7 Confidence interval

In this study, we will see how the central limit theorem comes in helps in constructing confidence intervals in any complicated situation. We have considered a binomial distribution and the confidence intervals are constructed by the two approaches,

- Exact distribution approach
- Asymptotic distribution approach

Let X be a discrete random variable where $X \sim Bin(n, p)$. Where 'n'(known) is the number of independent Bernoulli trials with probability of success 'p'(unknown) in each trial. The probability mass function of X is given by:

$$f(x) = \binom{n}{x} p^x (1-p)^{(n-x)}, \quad x = 0(1)n;\ p \in (0,1), n \in \mathbb{N}$$

7.1 Methodology

In this study, first, we will construct $100(1-\alpha)\%$ confidence intervals for p through both the approaches, where α is chosen to be 0.05. Now, we know that the confidence coefficient $(1-\alpha)$ which is 0.95 in this case, means that, if we draw repeated samples from the population, around 95% of the confidence intervals constructed from those samples, will capture the true unknown parameter 'p' on an average. So, considering the values of n as $20, 30, \ldots, 300$, we will be drawing 1000 random samples from $Bin(n, p)$ and then compute the empirical coverage for each choice of n. This process will be repeated for both approaches and we will compare the empirical coverages.

While drawing the random sample from $Bin(n, p)$, we will consider the value of p as 0.5.

7.2 Approach 1: Exact distribution

Let $F(\cdot)$ be the cumulative distribution function of X. Since X is a discrete random variable, it will be difficult to obtain a confidence interval for p explicitly, so we consider the following relationship between the cumulative distribution function of the binomial distribution and the regularized incomplete beta function, given by:

$$\begin{aligned} F(k; n, p) &= P(X \leq k) \\ &= \sum_{i=0}^{[k]} \binom{n}{i} p^i (1-p)^{(n-i)} \\ &= \frac{1}{B(n-k, k+1)} \int_0^{1-p} t^{n-k-1}(1-t)^k dt \\ &= I_{1-p}(n-k, k+1) \qquad \ldots\ldots (7) \end{aligned}$$

Now, if 'L' & 'U' be the lower and upper confidence limit of 'p', then the confidence interval for 'p' with confidence coefficient α is given by:

$$P(L \leq p \leq U) = 1 - \alpha$$

$$\Rightarrow F(U; n, p) - F(L; n, p) = 1 - \alpha \qquad \ldots\ldots (8)$$

Where, L & U are such that,

$$\frac{1}{B(x, n-x+1)} \int_{t=0}^{L} t^{x-1}(1-t)^{n-x} dt = \frac{\alpha}{2} \quad \ldots\ldots (9)$$

$$\frac{1}{B(x+1, n-x)} \int_{t=0}^{U} t^{x}(1-t)^{n-x-1} dt = 1 - \frac{\alpha}{2} \quad \ldots\ldots (10)$$

By solving the equations (9) & (10) numerically for various choices of n, we will get the confidence intervals (L, U) for p.

7.3 Approach 2: Asymptotic distribution

In this approach, we will proceed with the **Central Limit Theorem**. In case of $Bin(n, p)$, from *De Moivre-Laplace* central limit theorem, we have,

$$\frac{X - np}{\sqrt{np(1-p)}} \xrightarrow{D} N(0,1)$$

$$\Rightarrow \frac{\sqrt{n}(X/n - p)}{\sqrt{p(1-p)}} \xrightarrow{D} N(0,1), \quad \text{for large n}$$

$\therefore$ From the property of standard normal distribution, we have,

$$P\Big(-\tau_{\alpha/2} \leq \frac{\sqrt{n}(X/n - p)}{\sqrt{p(1-p)}} \leq \tau_{\alpha/2} \Big) = 1 - \alpha$$

$$\Rightarrow P\Big(\frac{X}{n} - \tau_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \leq p \leq \frac{X}{n} + \tau_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \Big) = 1 - \alpha$$

Since there is p in the expressions of limits, we estimate it by the sample proportion, $\frac{X}{n}$ and we get the asymptotic confidence interval with confidence coefficient α as:

$$P\Big(\frac{X}{n} - \tau_{\alpha/2} \sqrt{\frac{X(n-X)}{n^3}} \leq p \leq \frac{X}{n} + \tau_{\alpha/2} \sqrt{\frac{X(n-X)}{n^3}} \Big) = 1 - \alpha \quad \ldots\ldots (11)$$

Now, according to the abovesaid methodology, we have calculated the empirical coverages and it is plotted below to get an idea about how the central limit theorem comes in help over the exact approach for constructing confidence intervals in difficult situations.

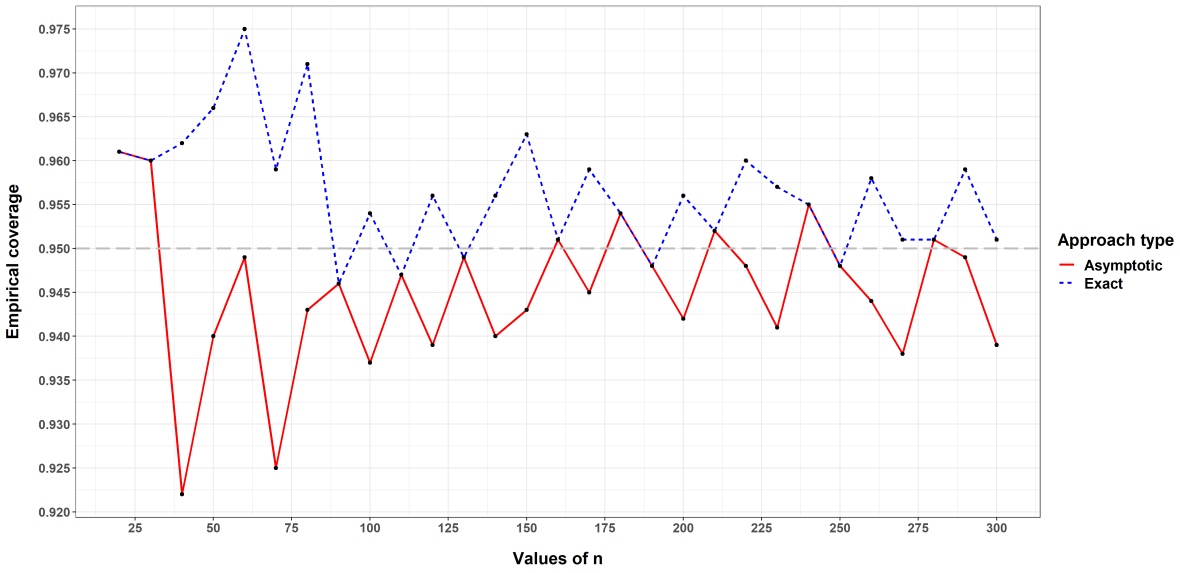

Figure 23

7.4 Conclusion

Now, from the above plot, it is clear that as the value of n is increasing, the empirical coverages obtained by both methods are getting closer to each other, and also it is quite clear that the way to derive a confidence interval through the 'Exact distribution' approach is very tedious, but through the application of central limit theorem, one can easily derive a $100(1-\alpha)\%$ confidence interval for the unknown binomial proportion p. So, this approach can be applied in many situations where it is laborious to derive the confidence intervals.

---

8 CLT for Dependent Random Variables

In order to work with the central limit theorem so far, we have considered a sequence of random variables which are independent and identically distributed but these conditions do not always need to hold true. It may be the case that the sequence of random variables is dependent on each other. In such cases, there may be difficulty to work with the central limit theorem often. Here we will illustrate how CLT can work for the dependent random variables with a suitable example.

Let $X_1, X_2, \ldots, X_n$ be a sequence of n dependent random variables such that,

$$E(X) = \theta \qquad Var(X) = \sigma^2 \qquad cov(X_i, X_j) \neq 0;\ \forall i \neq j$$

$$\therefore Var\{\sqrt{n}(\bar{X}_n - \theta)\} = \sigma^2 + \frac{1}{n}\sum\sum_{i \neq j} cov(X_i, X_j) \ldots (*)$$

Now, σ^2 is finite, so to assure the distribution convergence of $\sqrt{n}(\bar{X} - \theta)$, the second term in $(*)$ need to go to a finite limit. That is,

$$\lim_{n\to\infty} \frac{1}{n}\sum\sum_{i \neq j} cov(X_i, X_j) = \gamma \ldots (**)$$

$$\Rightarrow \lim_{n\to\infty} Var\{\sqrt{n}(\bar{X} - \theta)\} = \sigma^2 + \gamma$$

Thus, the central limit theorem in this case is defined as,

$$\frac{\sqrt{n}(\bar{X}_n - \theta)}{\sqrt{\sigma^2 + \gamma}} \xrightarrow{D} N(0, 1) \quad \text{for large n}$$

Remark: In the summation term of $(*)$, we have total $n(n-1)$ terms. So $(**)$ can occur in two ways,

- $cov(X_i, X_j) = 0$ for all but $n(n-1)$ number of pairs (i, j).
- $cov(X_i, X_j)$ tends to 0 sufficiently fast as $|i - j| \to 0$. So, the convergence of the sample means $\bar{X}_n$ will be different from the independent case depending upon the sign of γ.
    1. If $\gamma > 0$, then the convergence of sample mean towards normality will be slower than the case when the variables are independent.
    2. If $\gamma < 0$, the convergence will be faster than the independent case.

We now perform a simulation study to illustrate this.

8.1 Methodology

We will consider a multivariate normal distribution since here correlation between the random variables denotes that the random variables are dependent. We will draw a random vector of order n from $\mathbf{N}(\mu = \mathbf{0}, \mathbf{\Sigma})$ distribution, where the diagonal elements of $\mathbf{\Sigma}$ are 1 and for a fix i, $cov(X_i, X_j) \to 0$ as j differs from i. Here, the n dimensional random vector will be considered as a sequence of dependent observations $i.e.$ a sample of size n. Then we calculate the mean of the vector of observations. In this way, we generate $10,000$ such sample means and then we plot the histogram of the standardized mean values for various choices of sample size. The diagram is given below.

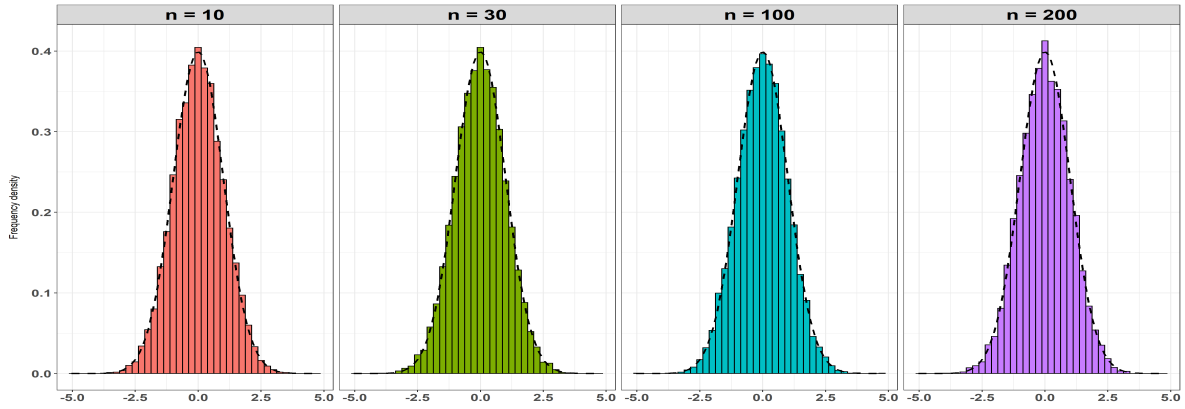

Figure 24

Summary table		
n	D_n	p-value
10	0.00642	0.80371
30	0.00749	0.62966
100	0.0059	0.87784
200	0.00532	0.93956

Table 5

8.2 Conclusion

From the diagram and also from the summary table, it is clear that normality is well achieved for all the cases which is also revealed by the p-values.

---

9 CLT for non-identical random variables

The classical central limit theorem deals with random variables which are independent and identical. But in many cases, the variables may not be identical. In this study, we will see how the central limit can be applied to the random variables which are independent but not necessarily identical.

Lyapounov proved the central limit theorem more generally with the assumption that the random variables have finite third-order moments. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution having CDF F_X. We have, $E(X_i) = \mu_i$ and $Var(X_i) = \sigma_i^2 \quad \forall\, i = 1(1)n$.

$$\text{Let } \sigma_n^2 = \sum_{i=1}^{n} Var(X_i) \quad \text{and let } Y = \sum_{i=1}^{n} X_i. \quad \text{If there exists } l > 0 \quad \text{such that,}$$

$$\lim_{n\to\infty} \left(\frac{1}{\sigma_n^{2+l}} \sum_{i=1}^{n} E\left[|X_i - \mu_i|^{2+l} \right] \right) = 0, \quad \text{then } Z = \frac{Y - E(Y)}{\sqrt{Var(Y)}} \xrightarrow{D} N(0,1).$$

$$\text{In particular,} \quad E(\bar{X}_n) = E\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_i$$

$$\text{and,} \quad Var(\bar{X}_n) = Var\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2,$$

9.1 Methodology

First we draw a random sample of size n from $N(\mu_i, \sigma_i^2)$ and then calculate the sample mean. We then standardize it by subtracting the average of the μ_i from the sample mean and dividing it by the square root of the average of σ_i^2; $i = 1(1)n$. This process is repeated $10,000$ times and then we plot the histograms of these observations for several choices of sample size n, in the separate panel of the same graph. The diagram is shown below.

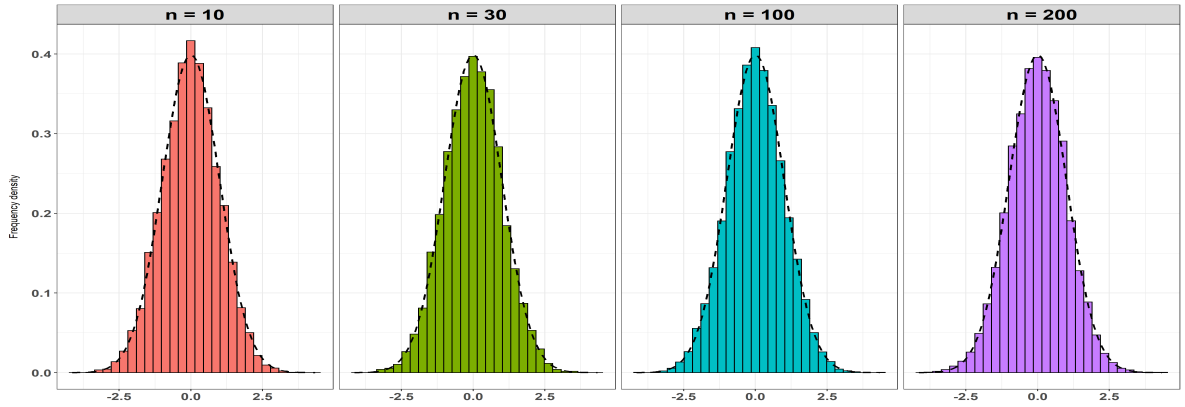

Figure 25

Summary table		
n	D_n	p-value
10	0.00834	0.49043
30	0.00693	0.72325
100	0.00473	0.97847
200	0.00677	0.74967

Table 6

9.2 Conclusion

From the diagram and also from the summary table, it is clear that normality is well achieved for all the cases which is also revealed by the p-values.

10 Conclusion

The central limit theorem is an important topic in the field of statistics. Some nice findings are revealed while doing this project. For example, although the limiting distribution of the sample mean is asymptotically normal but the asymptotic behaviour is different for different parent distributions, it depends upon the nature of the parent distribution.

It is very easy to apply CLT to real-life problems. Regardless of the size of the original population, we can always get an idea about the true population mean by considering a random sample of appropriate size.

CLT helps us a lot in handling many difficult inferential problems, under the consideration of large samples.

We have also seen that CLT can also be applied if one or more assumptions are not satisfied. For example, we can also apply CLT to the random variables which are not identical or dependent, by following some conditions.

11 Appendix

11.1 Kolmogorov-Smirnov test

Let X be a random variable with continuous distribution function F_X. To test,

$$H_0 : F_X(x) = F_0(x) \quad \forall x$$

$$F_X(x) \neq F_0(x) \quad \text{for some } x,$$

where $F_0(x)$ is completely specified distribution function.

Since the empirical distribution function $S_n(x)$ is the statistical image of the population distribution function $F_X(x)$, the difference between $S_n(x)$ and $F_0(x)$ should be very small except for sampling variation if the null hypothesis is true. Thus the Kolmogorov-Smirnov one-sample statistic is given by,

$$D_n = Sup_x|S_n(x) - F_0(x)|$$

Where D_n is a distribution free statistic under H_0. $\therefore$ we reject H_0 at α level of significance if $D_n > D_{n,\alpha}$, where $D_{n,\alpha}$ is such that $P(D_n > D_{n,\alpha}) = \alpha$.

11.2 Asymptotic distribution of Sample Quantiles

Let us suppose that $X_1, X_2, \ldots, X_n$ are $i.i.d.$ continuous random sample from a distribution having CDF $F_X(\cdot)$ and ξ_p be the p^{th} quantile of the distribution. Let us define a random variable Z_i in this way,

$$Z_i(\xi_p) = \begin{cases} 1, & if\ X \leq \xi_p \\ 0, & \text{otherwise} \end{cases}$$

And a random variable defined for fixed $\xi_p \in \mathbb{R}$ by:

$$Y_n(\xi_p) = \frac{1}{n}\sum_{i=1}^{n} Z_i \quad (\text{mean of } Z_i's)$$

Now, clearly here $Z_i's$ are Bernoulli random variable with probability of success $P(X_i \leq \xi_p) = F_X(\xi_p) = p$. So we have,

$$E(Z_i) = F_X(\xi_p) = p \quad \text{and} \quad Var(Z_i) = F_X(\xi_p)\{1 - F_X(\xi_p)\} = p(1-p)$$

Thus by the central limit theorem, we have,

$$\sqrt{n}(Y_n(\xi_p) - p) \overset{a}{\sim} N(0, p(1-p)) \quad \text{for large n}$$

Now let us consider a transformation through the function $g(t)$ $(0 < t < 1)$ by $g(t) = F_X^{-1}(t)$, and the first derivative of g is given by,

$$g^{(1)}(t) = \frac{d}{dt}F_X^{-1}(t) = \frac{1}{f_X(F_X^{-1}(t))}$$

$$\Big[\text{Since, } y = F_X^{-1}(t) \Leftrightarrow F_X(y) = t \Rightarrow f_X(y)dy = dt \Rightarrow \frac{dy}{dt} = \frac{1}{f_X(y)} = \frac{1}{f_X(F_X^{-1}(t))}\Big]$$

Then, by using **Delta method**, we get,

$$\sqrt{n}\Big(F_X^{-1}(Y_n(\xi_p)) - F_X^{-1}(p)\Big) \overset{a}{\sim} N\Big(0, \frac{p(1-p)}{\{f_X(F_X^{-1}(p))\}^2}\Big)$$

$$\Rightarrow \sqrt{n}\Big(F_X^{-1}(Y_n(\xi_p)) - \xi_p\Big) \overset{a}{\sim} N\Big(0, \frac{p\{1-p\}}{\{f_X(\xi_p)\}^2}\Big)$$

Now, $F_X^{-1}(Y_n(\xi_p))$ is a random variable that lies between the $(p-1)^{st}$ and p^{th} sample quantile, that can be written in terms of order statistic as $X_{(np)}$.

Therefore,

$$\sqrt{n}\Big(X_{(np)} - \xi_p\Big) \overset{a}{\sim} N\Big(0, \frac{p(1-p)}{\{f_X(\xi_p)\}^2}\Big)$$

11.3 R Codes

```r
library(tidyverse)
library(nleqslv)
library(MASS)
##=========================
## Code 1: Simulation study
##=========================

## Exponential distribution
exp_sim = function(n, t){
   df1 = 0  # data store
   df2 = matrix(ncol = 3, nrow = length(n))  # D max store
   R = 10000

   for(i in 1:length(n)){
      M = matrix(rexp(R*n[i], rate = 1/t), nrow = R, byrow = T)
      Z = sqrt(n[i])*(apply(M, 1, mean) - t)/t
      k = ks.test(Z, 'pnorm')
      df2[i,] = c(n[i], round(k$statistic,5),
                  round(k$p.value, 5))
      df1 = df1 %>% bind_cols(Z)
   }
   df1 = df1[,-1]
   colnames(df1) = paste('n =', n)

   df1 %>% pivot_longer(everything(),
                        names_to = 'sample_size',
                        values_to = 'x') %>%
      mutate(sample_size = factor(sample_size,
                                  levels = paste('n =',n))) %>%
      ggplot(aes(x = x, fill = sample_size)) +
      geom_histogram(aes(y = ..density..),
                     bins = 30, color = 'black') +
      stat_function(fun = dnorm, args = list(0,1),
                    lty = 2, lwd = 1) +
      facet_grid(~sample_size) +
      labs(x = '', y = 'Frequency density') +
      theme_bw() +
      theme(plot.margin = unit(c(1.2,0.8,1.2,0.8), "cm"),
            axis.title.y = element_text(vjust = 5),
            axis.text = element_text(face = 'bold', size = 14),
            legend.position = "none",
            strip.text = element_text(face = 'bold',
                                      colour = 'black',
                                      size = 20)) -> p

   rownames(df2) = paste(1:length(n))
   colnames(df2) = c('n','D_max','P value')
   return(list(df2,p))
}

exp_sim(c(10,30,100,200), 20)
setwd('D:/CLT Project')
ggsave(height = 9, width = 16,
       filename = 'Exp_sim(5).png')

## SIMILARLY, THE CODE FOR SIMULATION STUDY IN CASE OF OHTER DISTRIBUTIONS IS ALSO SAME.
```

```r
##==============================
## Code 2: Real-life application
##==============================
bw = read.csv("D:/Datasets/US_Baby_weights/us_new_born_2million.csv")
View(bw)

d = bw %>% filter(state == 'FL') %>%
   select(weight_pounds) %>% na.omit()
colnames(d) = c('wt')

sum(is.na(d))
summary(d)
dim(d)
head(d,10)
pm = mean(d$wt); pm

## Distribution of weights of babies:
ggplot(d, aes(wt)) + geom_histogram(bins = 40,
                                    colour = 'black',
                                    fill = 7,
                                    aes(y = ..density..)) +
   theme_minimal() +
   labs(title = 'Distribution of baby weights in the state',
        x = 'Baby weights (lbs)', y = 'Frequency density') +
   theme(plot.margin = unit(c(1,2,1,2), 'cm'),
         plot.title = element_text(face = 'bold',
                                   size = 28, hjust = 0.5),
         axis.title = element_text(face = 'bold', size = 20),
         axis.text = element_text(face = 'bold', size = 16),
         axis.title.x = element_text(vjust = -3),
         axis.title.y = element_text(vjust = 5)) +
   scale_x_continuous(n.breaks = 8) +
   geom_vline(xintercept = pm, colour = 'red', lty = 2,
              lwd = 1.5)

setwd("D:/CLT Project")
ggsave(filename = "population_dist.png",
       height = 9, width = 16)

## Distribution of sample means byt considering the sample of size 10:
R = 10000
n = 10
s1 = 0

for(i in 1:R)
   s1[i] = mean(sample(d$wt, replace = T, size = 10))

m1 = mean(s1); m1
s1 = as.data.frame(s1)
sd(s1$s1)

ggplot(s1, aes(s1)) + geom_histogram(bins = 40,
                                    colour = 'black',
                                    fill = 'lightblue',
                                    aes(y = ..density..)) +
   theme_minimal() +
   labs(title = 'n = 10',
        x = 'Sample means (lbs)', y = 'Frequency density') +
   theme(plot.title = element_text(face = 'bold',
```

```
                                    size = 20, hjust = 0.5),
        plot.margin = unit(c(1,2,1,2), 'cm'),
        axis.title = element_text(face = 'bold', size = 20),
        axis.text = element_text(face = 'bold', size = 16),
        axis.title.x = element_text(vjust = -3),
        axis.title.y = element_text(vjust = 5)) +
   scale_x_continuous(n.breaks = 8) +
   geom_vline(xintercept = c(pm,m1), colour = c('red','darkgreen'),
              lty = c(2,1))

setwd("D:/CLT Project")
ggsave(filename = "dist1.png",
       height = 9, width = 16)

## Distribution of sample means by considering successively higher sizes of samples:
s2 = 0
n = c(20,30,50,100,200,500)
R = 10000

for(i in 1:length(n))
   for(j in 1:R)
      s2[(i-1)*R+j] = mean(sample(d$wt, size = n[i], replace = T))


s3 = data.frame('x' = s2, 'n' = rep(n, each = R))
s3$n = as.factor(s3$n)
View(s3)

a = aggregate(x ~ n, mean, data = s3); a
colnames(a) = c('Sample size', 'means')

name = c('20' = 'n = 20',
         '30' = 'n = 30',
         '50' = 'n = 50',
         '100' = 'n = 100',
         '200' = 'n = 200',
         '500' = 'n = 500')

ggplot(s3, aes(x, fill = n)) +
   geom_histogram(aes(y = ..density..),
                  bins = 40, colour = 'black') +
   facet_wrap(.~n,
     labeller = labeller(n = as_labeller(name, label_context))) +
   labs(x = 'Sample means', y = 'Frequency density') +
   theme_bw() +
   theme(plot.margin = unit(c(2,0.8,2,0.8), "cm"),
         axis.title = element_text(face = 'bold', size = 20),
         axis.title.x = element_text(vjust = -3),
         axis.title.y = element_text(vjust = 5),
         axis.text = element_text(face = 'bold', size = 14),
         legend.position = "none",
         strip.text = element_text(face = 'bold',
                                   colour = 'black', size = 20))+
   scale_y_continuous(n.breaks = 8)

setwd("D:/CLT Project")
ggsave(filename = "dist2.png",
       height = 9, width = 16)
```

```r
##============================
## Code 3: Hypothesis testing
##============================
set.seed(0)

X = rcauchy(161,0.2,1)
Xm = median(X); Xm
k = 80

pdf = function(x){
   k_fact = factorial(2*k + 1)/(factorial(k)^2)
   ((0.5 + (1/pi)*atan(x))^k)*
      ((0.5 - (1/pi)*atan(x))^k)/(pi*(1+x^2))*k_fact
}

f = function(c) (c(integrate(pdf, c, Inf)$value - 0.05))

c_val = nleqslv(0, f)$x; c_val

# Power curve :
f = function(mu){   # power function

   k_fact = factorial(2*k + 1)/(factorial(k)^2)

   f = function(x){
      ((0.5 + (1/pi)*atan(x-mu))^k)*
         ((0.5 - (1/pi)*atan(x-mu))^k)/(pi*(1+(x-mu)^2))*k_fact
   }
   I = integrate(f, 0.205, Inf)$value
   return(I)
}

mu = seq(0,0.7, 0.001)
p1 = 0

## Storing the power values in both approaches:
for(i in 1:length(mu)) (p1[i] = f(mu[i]))
p2 = pnorm(8.077799*mu - 1.644854)

df = data.frame('mu' = rep(mu,2),
                'power' = c(p1,p2),
                'Type' = rep(c('Exact distribution',
                               'Asymptotic distribution'),
                             each = length(mu)))

my_col = c('red', 'blue')
df %>% ggplot(aes(x = mu, y = power, colour = Type)) +
   geom_line(aes(colour = Type, linetype = Type),
             lwd = 2, alpha = 0.4) +
   theme_minimal() +
   labs(x = 'Parameter values',
        y = 'Power values')+
   theme(axis.title = element_text(face = 'bold',
                                   size = 18),
         plot.margin = unit(c(0.7,0.7,0.7,1.2), 'cm'),
         legend.title = element_text(face = 'bold',
                                     size = 16),
         legend.text = element_text(face = 'bold',
                                    size = 14),
```

```
        axis.text = element_text(face = 'bold',
                                 size = 14),
        axis.title.x = element_text(vjust = -3),
        axis.title.y = element_text(vjust = 5)) +
   scale_x_continuous(n.breaks = 12) +
   scale_y_continuous(n.breaks = 14, minor_breaks = NULL) +
   geom_hline(yintercept = 0, lty = 2, lwd = 1) +
   geom_vline(xintercept = 0, lty = 2, lwd = 1) +
   scale_color_manual(values = my_col)

setwd("D:/CLT Project")
ggsave(filename = 'power curves.png',
       height = 9, width = 16)
```

```
##============================
## Code 4: Confidence interval
##============================

## Function for solving the confidence limits:
CI_func = function(n,x)
{
   a = 0.05 ## level of significance

   fl = function(l){
      f = function(t){
         (t^(x-1))*((1-t)^(n-x))
      }
         (integrate(f, 0, l)$value)/beta(x, n-x+1) - a/2
   }
   c1 = nleqslv(x/n, fl)$x

   fu = function(u){
      f = function(t){
         (t^x)*((1-t)^(n-x-1))
      }
         (integrate(f, 0, u)$value)/beta(x+1, n-x) - (1-a/2)
   }
   c2 = nleqslv(x/n, fu)$x

   return(c(c1,c2))
}

## Function for constructing empirical confidence coefficients:
coeff_func = function(n, condition){
   R = 1000; CI = 0
   count1 = 0; count2 = 0

   set.seed(1)
   for(i in 1:R){
      x = rbinom(1, n, 0.5)
      CI = CI_func(n, x)
      if(CI[1] < 0.5 & CI[2] > 0.5) (count1 = count1 + 1)
   }

   set.seed(1)
   for(i in 1:R){
      x = rbinom(1, n, 0.5)
      c1 = (x/n) - qnorm(0.975)*sqrt(x*(n-x)/(n^3))
      c2 = (x/n) + qnorm(0.975)*sqrt(x*(n-x)/(n^3))
```

```
      if(c1 < 0.5 & c2 > 0.5) (count2 = count2 + 1)
   }

   if(condition == 'exact'){
      return(count1/R)
   }else if(condition == 'asym'){
      return(count2/R)
   }
}

N = seq(20,300,10)
cov_exact = 0
cov_asym = 0

for(i in 1:length(N)){
   cov_exact[i] = coeff_func(N[i], 'exact')
   cov_asym[i] = coeff_func(N[i], 'asym')
}

df = tibble(N, 'Exact' = cov_exact,
      'Asymptotic' = cov_asym)

my_col = c('red','blue')
df %>% pivot_longer(Exact:Asymptotic,
                    names_to = 'Approach type',
                    values_to = 'coeff') %>%
   ggplot(aes(x = N, y = coeff, color = `Approach type`,
              linetype = `Approach type`)) +
   geom_line(lwd = 1) + geom_point(color = 'black')+
   theme_bw() +
   labs(x = 'Values of n', y = 'Empirical coverage') +
   theme(axis.title = element_text(face = 'bold', size = 16),
         axis.text = element_text(face = 'bold', size = 12),
         plot.margin = unit(c(1.5,0.7,3,1), 'cm'),
         axis.title.x = element_text(vjust = -3),
         axis.title.y = element_text(vjust = 5),
         legend.title = element_text(face = 'bold', size = 16),
         legend.text = element_text(face = 'bold', size = 14)) +
   scale_x_continuous(n.breaks = 12) +
   scale_y_continuous(n.breaks = 10) +
   scale_color_manual(values = my_col) +
   geom_hline(yintercept = 0.95, lwd = 1, color = 'grey',
              lty = 5)

setwd("D:/CLT Project")
ggsave(filename = "CI_plot(new).png",
       height = 9, width = 16)
```

```
##=========================================
## Code 5: CLT for dependent random variables
##=========================================

## Function to generate a covariance matrix of desired order:
cov_func = function(k){
  M = matrix(ncol = k, nrow = k)
  diag(M) = rep(1,k)

  for(i in 1:(k-1)){
    r = round(rev(sort(rexp(k-i, 15))),3)
```

```
    M[i,(i+1):k] = r
    M[(i+1):k,i] = r
  }
  return(M)
}
## Function to perform the simulation study:
corr_sim = function(n){
   R = 10000
   df1 = 0
   df2 = matrix(ncol = 3, nrow = length(n))

   for(i in 1:length(n)){
      cov_mat = cov_func(n[i])
      M = mvrnorm(10000, rep(0,n[i]), cov_mat)

      Z = sqrt(n[i])*apply(M, 1, mean)/sqrt(1+
                          (sum(cov_mat)-n[i])/n[i])
      k = ks.test(Z, 'pnorm')
      df2[i,] = c(n[i], round(k$statistic,5),
                  round(k$p.value, 5))
      df1 = df1 %>% bind_cols(Z)
   }
   df1 = df1[,-1]
   colnames(df1) = paste('n =', n)

   df1 %>% pivot_longer(everything(),
                        names_to = 'sample_size',
                        values_to = 'x') %>%
      mutate(sample_size = factor(sample_size,
                                  levels = paste('n =',n))) %>%
      ggplot(aes(x = x, fill = sample_size)) +
      geom_histogram(aes(y = ..density..),
                     bins = 40, color = 'black') +
      stat_function(fun = dnorm, args = list(0,1),
                    lty = 2, lwd = 1) +
      facet_grid(~sample_size) +
      labs(x = '', y = 'Frequency density') +
      theme_bw() +
      theme(plot.margin = unit(c(1.2,0.8,1.2,0.8), "cm"),
            axis.title.y = element_text(vjust = 5),
            axis.text = element_text(face = 'bold', size = 14),
            legend.position = "none",
            strip.text = element_text(face = 'bold',
                                      colour = 'black',
                                      size = 20)) -> p

   colnames(df2) = c('n','D_max','P value')
   rownames(df2) = paste(1:length(n))

   return(list(p,df2))
}

corr_sim(c(10,30,100,200))

setwd('D:/CLT Project')
ggsave(height = 9, width = 16,
       filename = 'corr.png')
```

```r
##=============================================
## Code 6: CLT for non-identical random variables
##=============================================
nonid_func = function(n){
   df1 = 0  # data store
   df2 = matrix(ncol = 3, nrow = length(n))  # D max store
   R = 10000
   set.seed(1)

   for(i in 1:length(n)){
      m = runif(n[i],0,10)
      sd = runif(n[i],1,3)
      M = matrix(ncol = n[i], nrow = R)
      for(j in 1:R) (M[j,] = rnorm(n[i], m, sd))
      Z = (apply(M, 1, mean) - mean(m))/sqrt(sum(sd^2)/(n[i]^2))
      k = ks.test(Z, 'pnorm')
      df2[i,] = c(n[i], round(k$statistic,5),
                  round(k$p.value, 5))
      df1 = df1 %>% bind_cols(Z)
   }

   df1 = df1[,-1]
   colnames(df1) = paste('n =', n)

   df1 %>% pivot_longer(everything(),
                        names_to = 'sample_size',
                        values_to = 'x') %>%
      mutate(sample_size = factor(sample_size,
                                  levels = paste('n =',n))) %>%
      ggplot(aes(x = x, fill = sample_size)) +
      geom_histogram(aes(y = ..density..),
                     bins = 30, color = 'black') +
      stat_function(fun = dnorm, args = list(0,1),
                    lty = 2, lwd = 1) +
      facet_grid(~sample_size) +
      labs(x = '', y = 'Frequency density') +
      theme_bw() +
      theme(plot.margin = unit(c(1.2,0.8,1.2,0.8), "cm"),
            axis.title.y = element_text(vjust = 5),
            axis.text = element_text(face = 'bold', size = 14),
            legend.position = "none",
            strip.text = element_text(face = 'bold',
                                      colour = 'black',
                                      size = 20)) -> p
   rownames(df2) = paste(1:length(n))
   colnames(df2) = c('n','D_max','P value')
   return(list(df2,p))
}

nonid_func(c(10,30,100,200))
setwd('D:/CLT Project')
ggsave(height = 9, width = 16,
       filename = 'nonid.png')
```

12 Acknowledgement

I would like to express my special gratitude to Father Principal Rev. Dr Dominic Savio, Sj. and the Department of Statistics who gave me the golden opportunity to do this wonderful project on 'Central Limit Theorem: Simulation and Application'. I would also like to thank my supervisor and dissertation guide Professor Madhura Das Gupta for her guidance throughout the duration of the completion of my project. I am also very much grateful to all my respected professors of St. Xavier's College Statistics faculty, who have inculcated in me, a strong research mindset, curiosity and intrinsic capability to pursue this subject. Lastly, I would like to thank St. Xavier's College for the opportunity to prepare a dissertation project paper on a topic of my choice, as well as for imbibing in me a drive to polish my research mindset.

13 References

1. Elements of Large-Sample Theory - *E.L. Lehmann* (Springer)
2. Bento, C. (2021, December 16). Central Limit Theorem: a real-life application - Towards Data Science. Medium. https://towardsdatascience.com/central-limit-theorem-a-real-life-application-f638657686e1
3. Central limit theorem. Wikipedia. https://en.wikipedia.org/wiki/Central_limit_theorem
4. Binomial proportion confidence interval. Wikipedia. https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval