# Mathematics for Engineers II
# Project Report

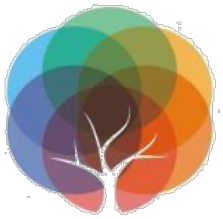
**PROJECT NO : 13**

A Project Submitted
in Partial Fulfilment of the Requirements for the Degree of
**Bachelor of Technology**
in
**Computer Science**

**Prepared**
**By**
**Group - 7**

Group Members :
Anshul Yadav (220643)
Shruti (220576)
Ritika Yadav (220592)

**Submitted**
**To**
**Dr. Rishi Asthana**



BML MUNJAL UNIVERSITY™

SCHOOL OF ENGINEERING AND TECHNOLOGY
BML MUNJAL UNIVERSITY GURGAON, 2023

## Content :

## **Problem Statement** :

Application of central limit theorem in tossing of a coin and birthday on any given day of the week separately. Perform the experiment and use the simulation in the software. Use R wherever required.

## Introduction :

The Central Limit Theorem (CLT) is the cornerstone of probability theory and statistics. This approach is effective because it uses a population's sample to shed light on the behaviour of the entire population. Whatever the population distribution, a normal distribution of the sample means would be reached in CLT as sample size increased.

In this project, we aim to demonstrate the application of the CLT in the context of coin tossing and birthday distribution. We will perform experiments by simulating a large number of coin tosses and generating random dates to represent birthdays. We will then use the R software to analyse the results and demonstrate the combining the sample means to the normal distribution.

## Mathematical Theory and Formula :

With a mean μ and a limited variance σ^2, let X be a random variable. Let x1, x2,..., xn be a random sample of size n that was taken from the population with the distribution of x in a consistent, independent manner. Then the sample mean would

$$\bar{X} = (X1 + X2 + \ldots + Xn)/n$$

have a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, as $n \to \infty$.

Assume that X is a random variable with mean μ and standard deviation σ. The CLT predicts that as sample size n approaches infinity, the distribution of the sample means ($\bar{X}$) will be roughly normal, with a mean and standard deviation and $\sigma/\sqrt{n}$.

In mathematical notation, we can represent this as follows:

$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$

where n is the sample size and $\bar{X}$ is the sample mean, μ population mean, and σ population standard deviation.

## Analysis (CLT) :

- **Central Limit Theorem -**

In statistics, the Central Limit Theorem (CLT), a concept from probability theory, has been widely applied. The underlying premise of the theory is that, regardless of the initial population distribution pattern, the distribution of the sample mean tends to follow the normal distribution as sample sizes for random variables rise. Additionally, CLT shows that when sample size grows and the sample has a normal distribution, the variation of the sample mean approaches the population variance.

Abraham de Moivre first proposed the CLT in 1733, but it wasn't officially established until the 1930s, when prominent Hungarian mathematician George Pólya popularised the phrase "central limit theorem." The theorem has had a substantial influence on the advancement of contemporary statistical theory and its applications since that time.

CLT is significant because it gives academics a way to estimate population parameters by allowing them to draw conclusions about the population from their sample. In disciplines like finance, economics, and the social sciences, where access to entire populations is frequently restricted to researchers, this is of enormous practical importance.

In essence, the CLT offers a method for approximating the behaviour of sample means from any population and gaining knowledge of its characteristics. It is a fundamental idea that has significantly advanced the area of statistics, making it a vital resource for both scholars and practitioners.

- **Working :**

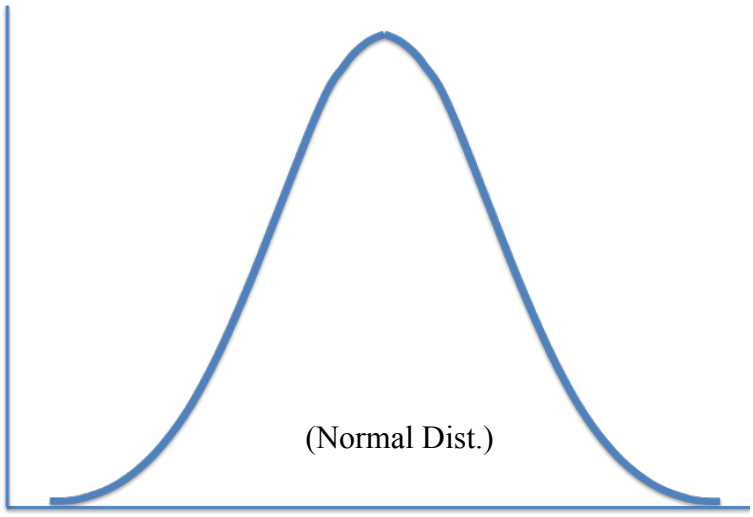
The CLT is a powerful concept in statistics, providing valuable insights into the behaviour of random variables. This theorem deals with selecting samples of size n from a population when the mean $\mu$ and standard deviation $\sigma$ are known.
According to the Central Limit Theorem, there are two rules that apply when collecting samples of size n.
The first rule asserts that if the sample size is "large enough" (equal to or larger than 30), the histogram of sample means will roughly resemble a typical bell shape. In other words, regardless of the underlying distribution of the individual variables, the distribution of the sample means will gravitate towards a normal distribution as the sample size grows.
The second rule indicates that a histogram will likewise have a typical bell shape if we gather samples of size n that are "large enough," add each sample, and produce a histogram. Since the sample means (averages) and sums will both have a normal distribution, precise population parameter estimation is attainable.
Overall, the CLT is a powerful and useful tool in statistics, allowing us to draw accurate inferences about population parameters based on limited samples. By understanding and utilising the CLT, we can gain valuable insights into the behaviour of random variables and improve our ability to analyse and interpret data.

- **Understanding -**

Central Limit Theorem (CLT) is a statistical theory, Regardless of how the data are actually distributed, the sample mean of the data will eventually equal the population mean as the sample size increases. This means that whether or not the population's distribution is normal has no impact on the data's accuracy.

Typically, a sample size of around 30-50 or greater is considered sufficient for the CLT to hold, resulting in a fairly normally distributed graph of the sample means. As more samples are taken, the distribution of the sample means increasingly resembles a normal distribution.

In summary, With the use of the CLT, which is an effective statistical method, it is possible to estimate the population mean and standard deviation using just sample data. CLT makes guarantee that the sample means gravitate towards a normal distribution as the sample size increases, improving the accuracy of statistical inference and prediction. The combination of the CLT with the law of large numbers provides a robust framework for analysing data and making predictions about populations.




- **Components -**

CLT consists of some important features. These features primarily revolve around sample, sample size, and data population.

1. **Sampling is continuous -** means that some pattern units will match previously selected pattern units.
2. **Sampling is random** - All the samples should be randomly selected to have similar statistical probability.
3. **Samples should be independent** - Future samples and the outcomes of additional samples shouldn't be impacted by decisions or results from one sample..
4. **Samples should be limited** - It is said that a sample should not exceed 10% of the population if it is taken without the sample substitution.
5. **Larger sample size** - CLT is used when a large sample size is selected.

**Simulations :**

CODE used to implement the simulation :

- To set file location :

| | | |
|---|---|---|
| setwd(“file location") | — | Setwd is to set the dictionary |
| mydata <- read.csv(“Name.csv") | — | read.csv is to read the csv input file. |
| tail (mydata) | — | Tail shows the last 5 values of the file. |

- To create a function of ’n’ sample size and ‘x’ random samples :

| | | |
|---|---|---|
| sample_func <- function { | — | Create a function |
| sample(mydata$Colname, size=n, replace=TRUE) | — | Take a sample from mydata file and using Column name , of size n |
| } | | |
| samples <- replicate(x, sample_func()) | — | Taking out the random x outcomes/ samples |

- To calculate the sum of each sample :

| | | |
|---|---|---|
| sample_sums <- rowSums (samples) | — | To get the sum of rows of samples |

- To plot the histogram :

| | | |
|---|---|---|
| hist (sample_sums, main-"Histogram", xlab="Sum of n") | — | To get the histogram graph to the sample_sums |

1. Application of central limit theorem in tossing of a coin for 500 cases to coin toss. Using R simulation to plot the frequency histogram of the coin toss rate.

(Input file is attached with report)

A. 100 samples of size 5

Code :

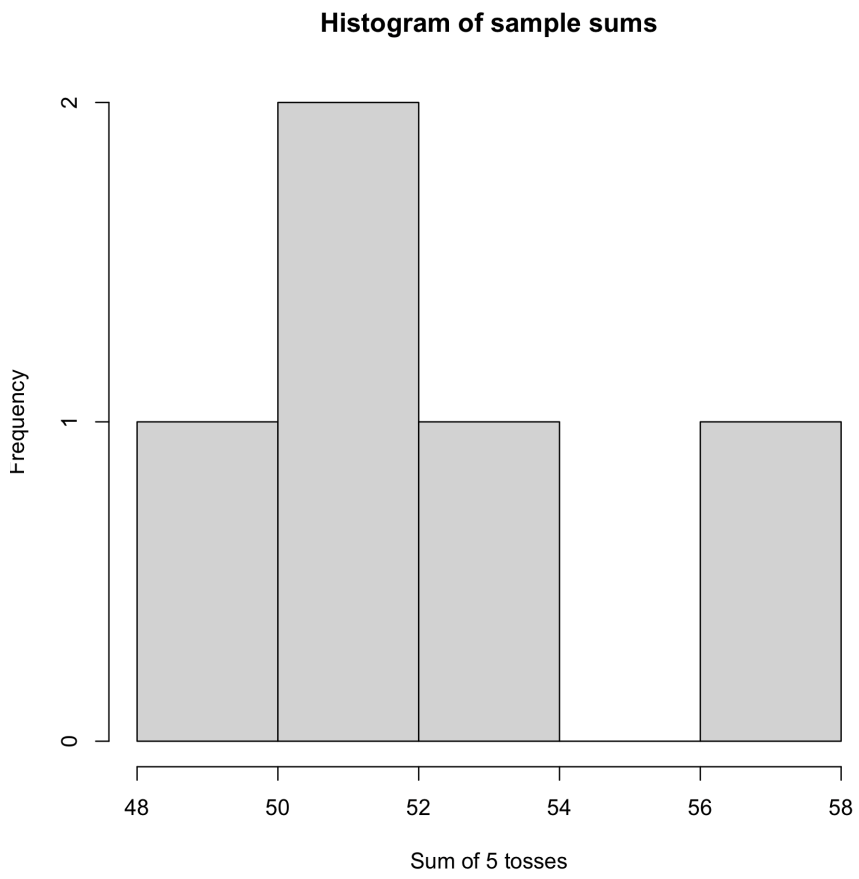
```
setwd("/Users/anshulyadav/Documents/")
cointoss <- read.csv("coin_toss.csv")
tail (cointoss)

sample_func <- function(){
sample(cointoss$Outcome, size=5, replace=TRUE)
}

samples <- replicate(100, sample_func())
sample_sums <- rowSums (samples)

hist (sample_sums, main-"Histogram of Sample Sums", xlab="Sum of 5 Tosses")
```

GRAPH-



B. 100 samples of size 10

Code :

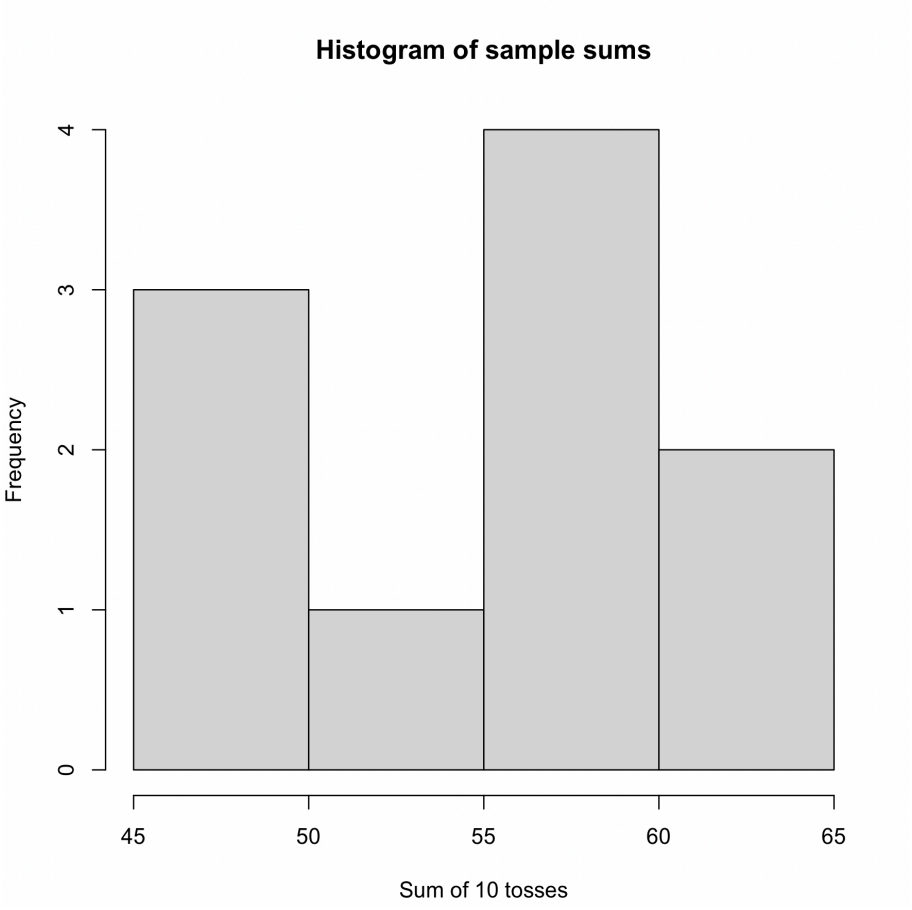
```
setwd("/Users/anshulyadav/Documents/")
cointoss <- read.csv("coin_toss.csv")
tail (cointoss)

sample_func <- function(){
sample(cointoss$Outcome, size=10, replace=TRUE)
}

samples <- replicate(100, sample_func())
sample_sums <- rowSums (samples)

hist (sample_sums, main-"Histogram of Sample Sums", xlab="Sum of 10 Tosses")
```

GRAPH-



C. 100 samples of size 15

Code :

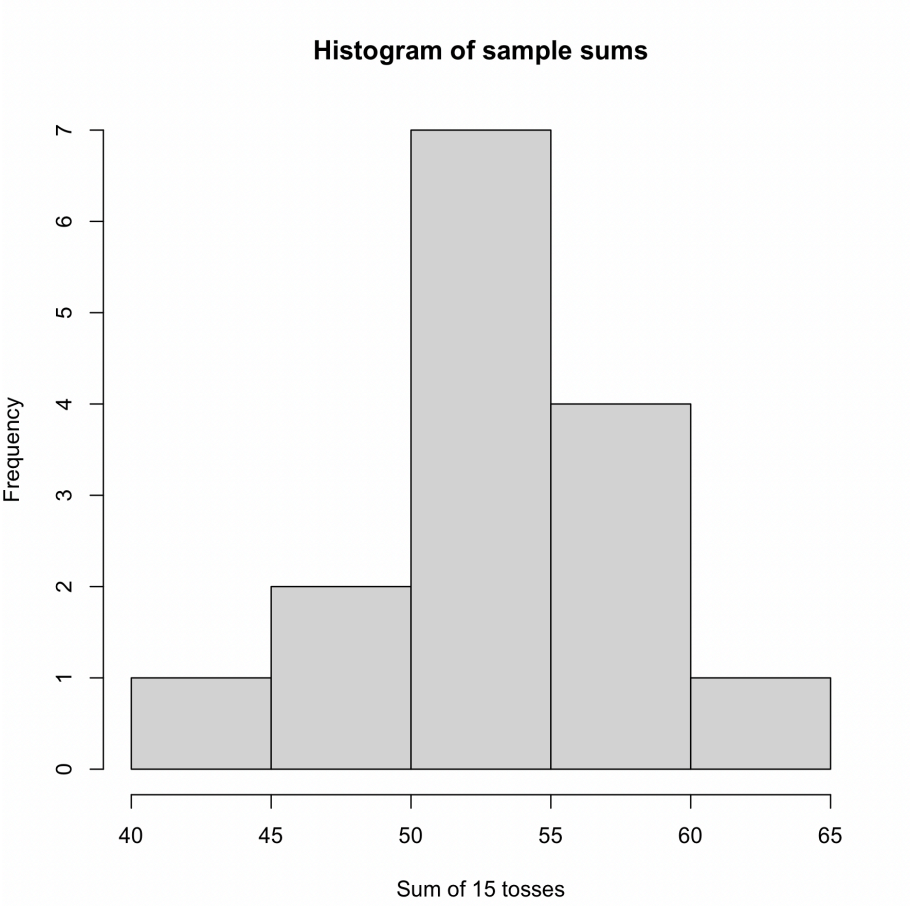
```
setwd(“/Users/anshulyadav/Documents/”)
cointoss <- read.csv(“coin_toss.csv”)
tail (cointoss)

sample_func <- function(){
sample(cointoss$Outcome, size=15, replace=TRUE)
}

samples <- replicate(100, sample_func())
sample_sums <- rowSums (samples)

hist (sample_sums, main-"Histogram of Sample Sums", xlab=“Sum of 15 Tosses”)
```

GRAPH-



D. 100 samples of size 30

Code :

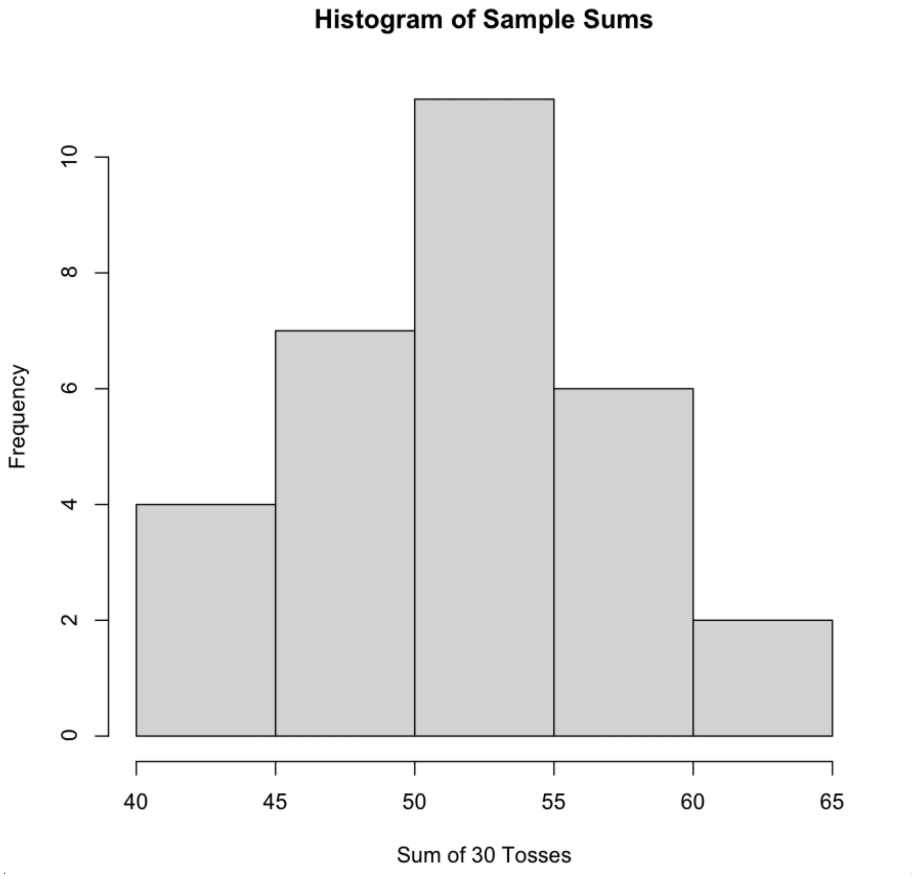
```
setwd("/Users/anshulyadav/Documents/")
cointoss <- read.csv("coin_toss.csv")
tail (cointoss)

sample_func <- function(){
sample(cointoss$Outcome, size=30, replace=TRUE)
}

samples <- replicate(100, sample_func())
sample_sums <- rowSums (samples)

hist (sample_sums, main-"Histogram of Sample Sums", xlab="Sum of 30 Tosses")
```

GRAPH-



E. 100 samples of size 50

Code :

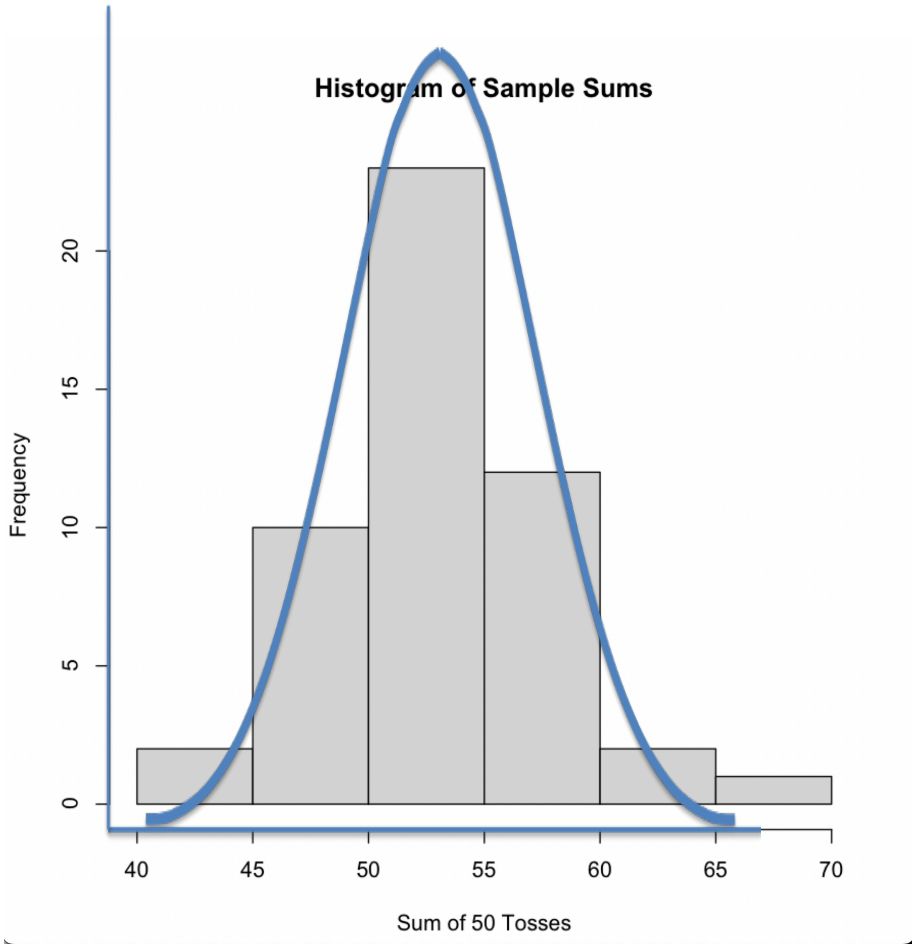
```
setwd("/Users/anshulyadav/Documents/")
cointoss <- read.csv("coin_toss.csv")
tail (cointoss)

sample_func <- function(){
sample(cointoss$Outcome, size=50, replace=TRUE)
}

samples <- replicate(100, sample_func())
sample_sums <- rowSums (samples)

hist (sample_sums, main-"Histogram of Sample Sums", xlab="Sum of 50 Tosses")
```

GRAPH-



F. 100 samples of size 80

Code :

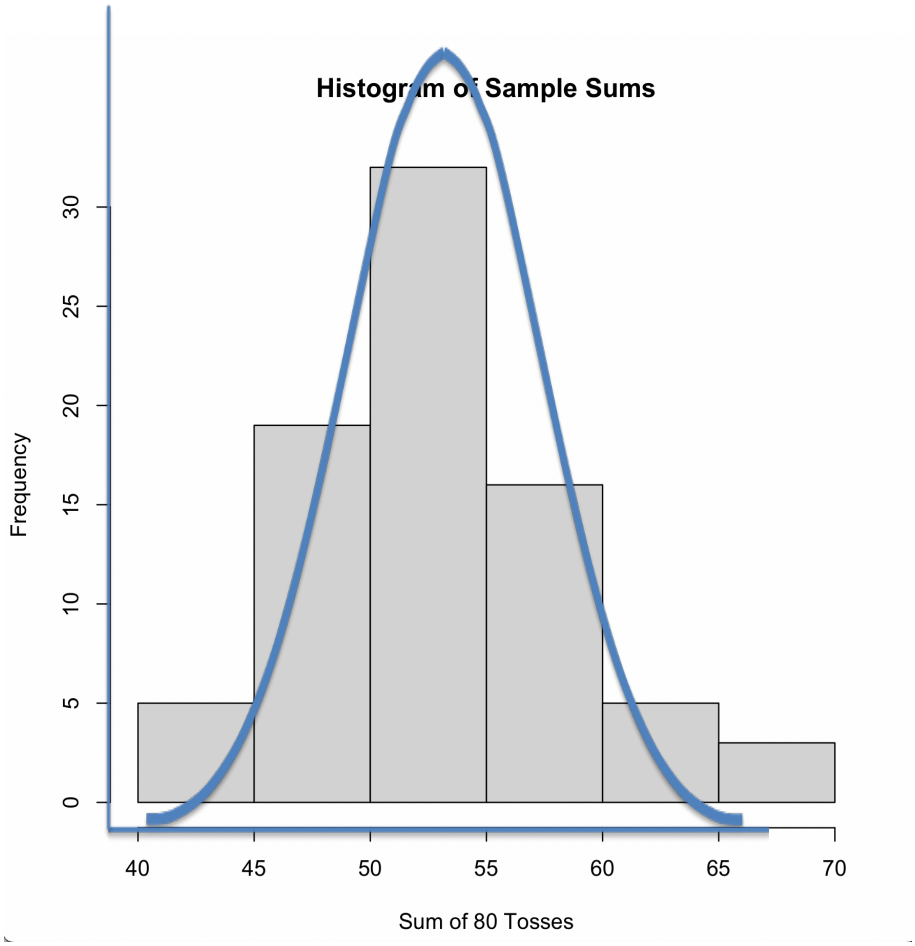
```
setwd("/Users/anshulyadav/Documents/")
cointoss <- read.csv("coin_toss.csv")
tail (cointoss)

sample_func <- function(){
sample(cointoss$Outcome, size=80, replace=TRUE)
}

samples <- replicate(100, sample_func())
sample_sums <- rowSums (samples)

hist (sample_sums, main-"Histogram of Sample Sums", xlab="Sum of 50 Tosses")
```

GRAPH-



G. 50 samples of size 30

Code :

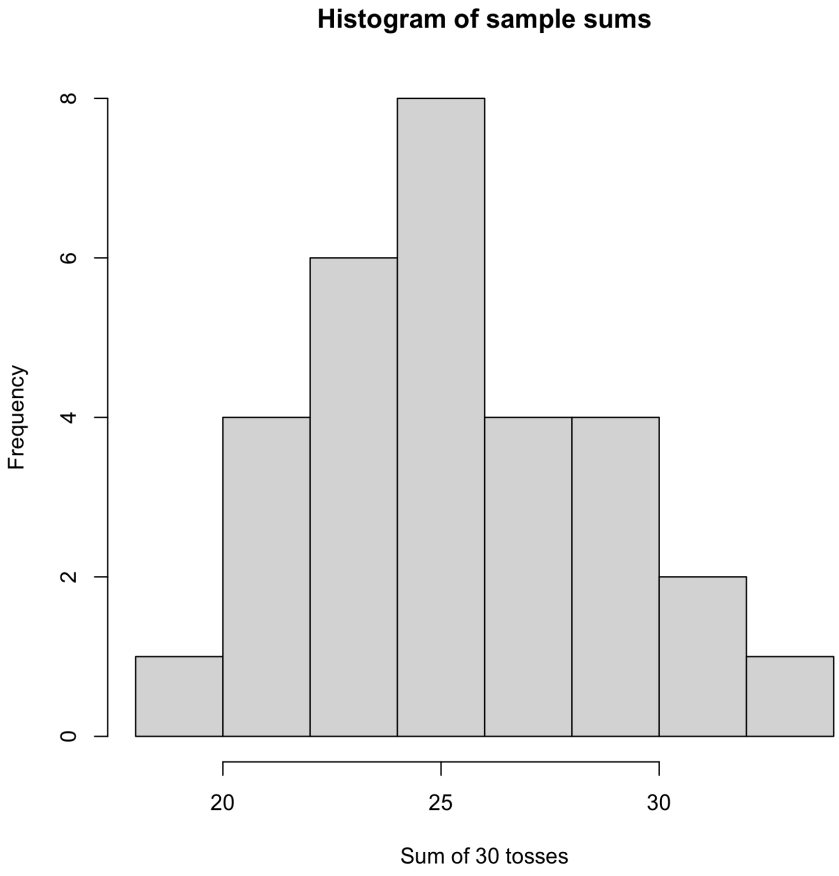
```
setwd("/Users/anshulyadav/Documents/")
cointoss <- read.csv("coin_toss.csv")
tail (cointoss)

sample_func <- function(){
sample(cointoss$Outcome, size=30, replace=TRUE)
}

samples <- replicate(50, sample_func())
sample_sums <- rowSums (samples)

hist (sample_sums, main-"Histogram of Sample Sums", xlab="Sum of 50 Tosses")
```

GRAPH-



H. 50 samples of size 50

Code :

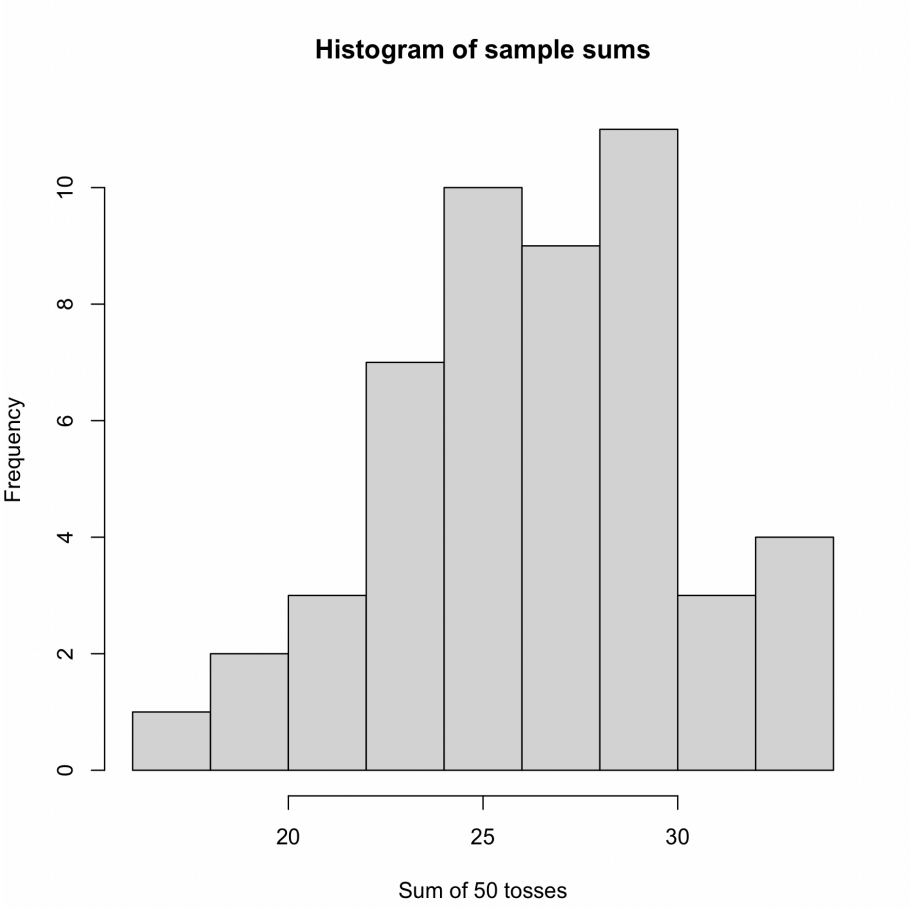
```
setwd("/Users/anshulyadav/Documents/")
cointoss <- read.csv("coin_toss.csv")
tail (cointoss)

sample_func <- function(){
sample(cointoss$Outcome, size=50, replace=TRUE)
}

samples <- replicate(50, sample_func())
sample_sums <- rowSums (samples)

hist (sample_sums, main-"Histogram of Sample Sums", xlab="Sum of 50 Tosses")
```

GRAPH-



I. 200 samples of size 30

Code :

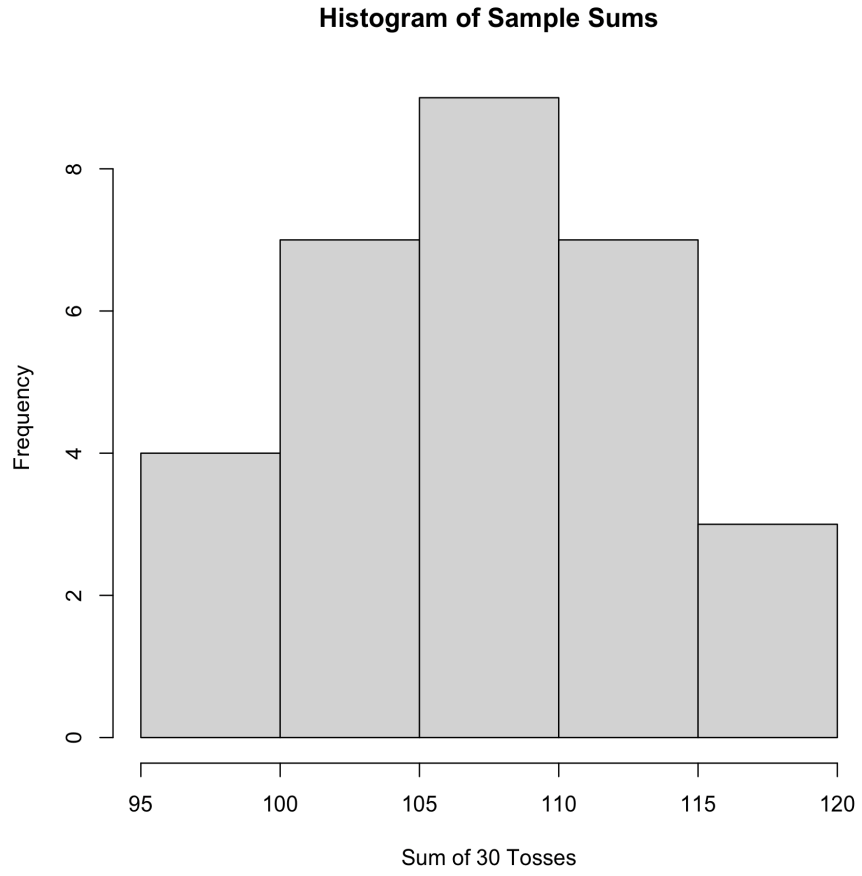
```
setwd("/Users/anshulyadav/Documents/")
cointoss <- read.csv("coin_toss.csv")
tail (cointoss)

sample_func <- function(){
sample(cointoss$Outcome, size=30, replace=TRUE)
}

samples <- replicate(200, sample_func())
sample_sums <- rowSums (samples)

hist (sample_sums, main-"Histogram of Sample Sums", xlab="Sum of 30 Tosses")
```

GRAPH-



J. 200 samples of size 50

Code :

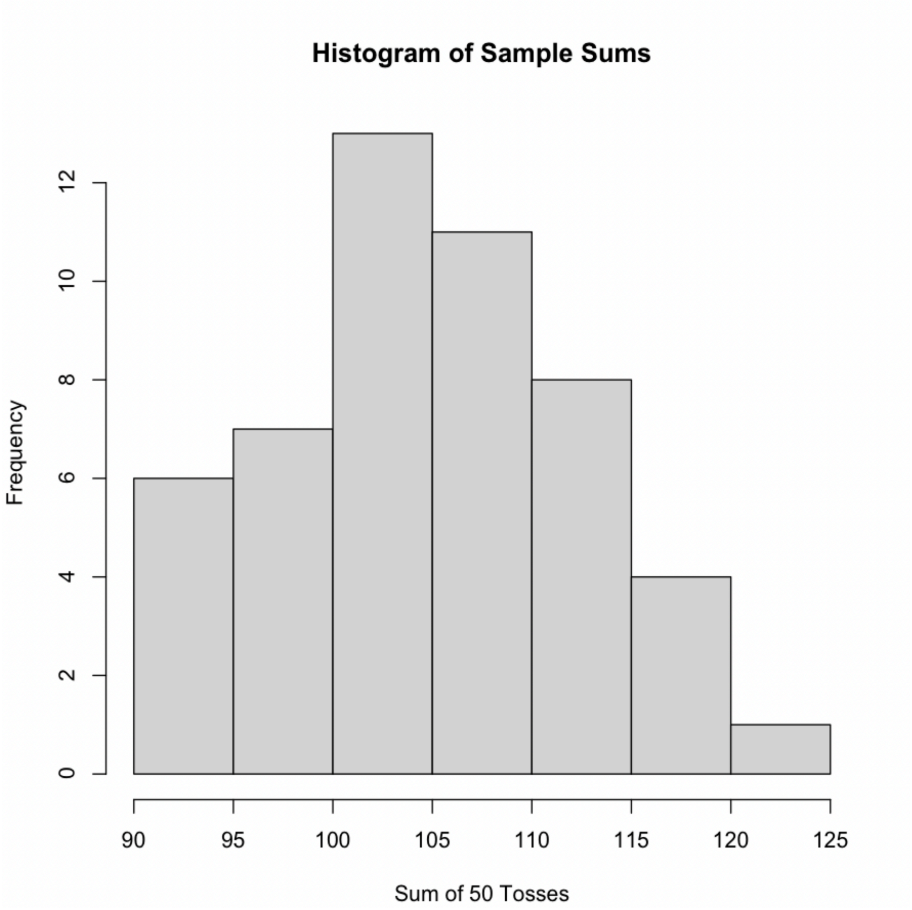
```
setwd("/Users/anshulyadav/Documents/")
cointoss <- read.csv("coin_toss.csv")
tail (cointoss)

sample_func <- function(){
sample(cointoss$Outcome, size=50, replace=TRUE)
}

samples <- replicate(200, sample_func())
sample_sums <- rowSums (samples)

hist (sample_sums, main-"Histogram of Sample Sums", xlab="Sum of 50 Tosses")
```

GRAPH-



2. Application of central limit theorem in birthday of IPL player on any given day of the week. Use numbers 1,2,3 etc for Sun, Mon, Tuesday etc while plotting. Plot the frequency histogram of the birthday, using R simulation.

(Input file is attached with report)

I. 100 samples of size 5

Code :

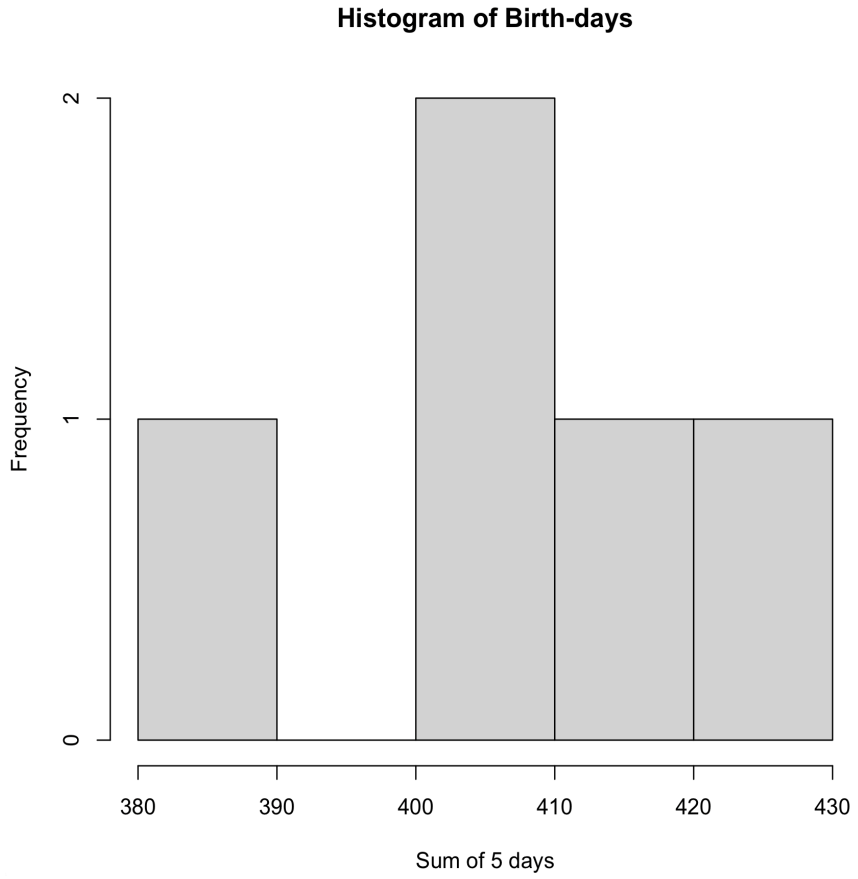
```
setwd("/Users/anshulyadav/Documents/")
ipl<- read.csv("iplbd.csv")
tail (cointoss)

sample_func <- function(){
sample(ipl$Birthday, size=5, replace=TRUE)
}

samples <- replicate(100, sample_func())
sample_sums <- rowSums (samples)

hist(sample_sums, main-"Histogram of Birth-days", xlab="Sum of 5 days")
```

GRAPH-



II. 100 samples of size 10

Code :

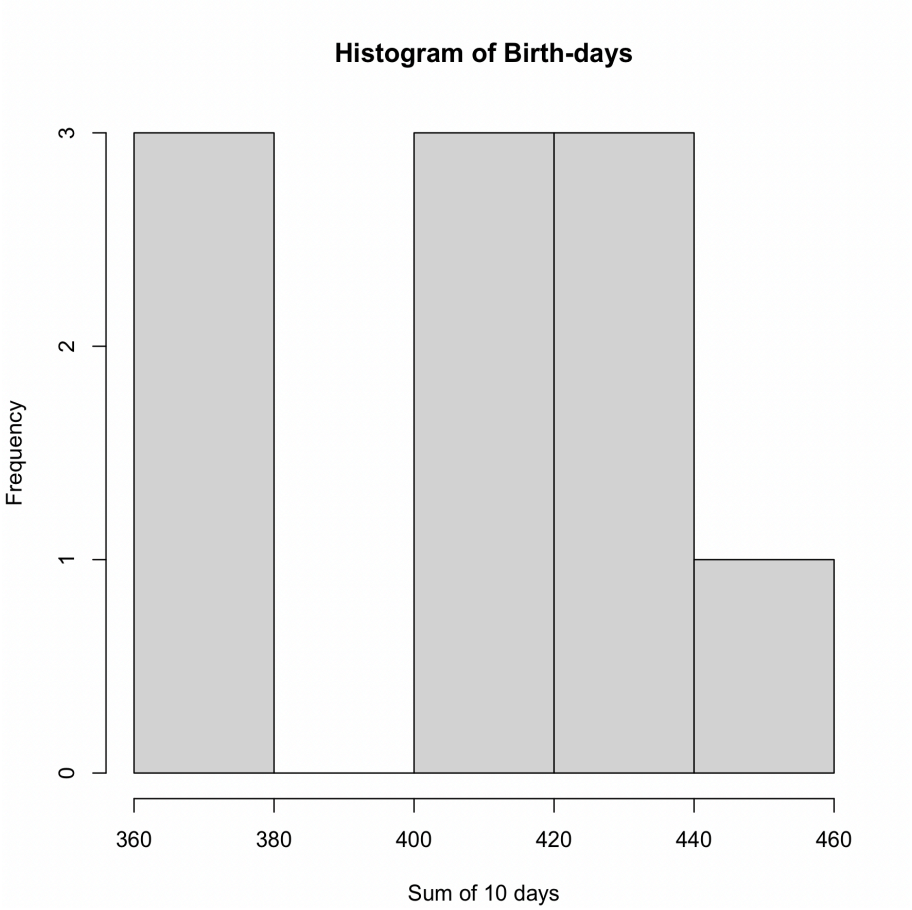
```
setwd("/Users/anshulyadav/Documents/")
ipl<- read.csv("iplbd.csv")

sample_func <- function(){
sample(ipl$Birthday, size=10, replace=TRUE)
}

samples <- replicate(100, sample_func())
sample_sums <- rowSums (samples)

hist(sample_sums, main-"Histogram of Birth-days", xlab="Sum of 10 days")
```

GRAPH-



III. 100 samples of size 15

Code :

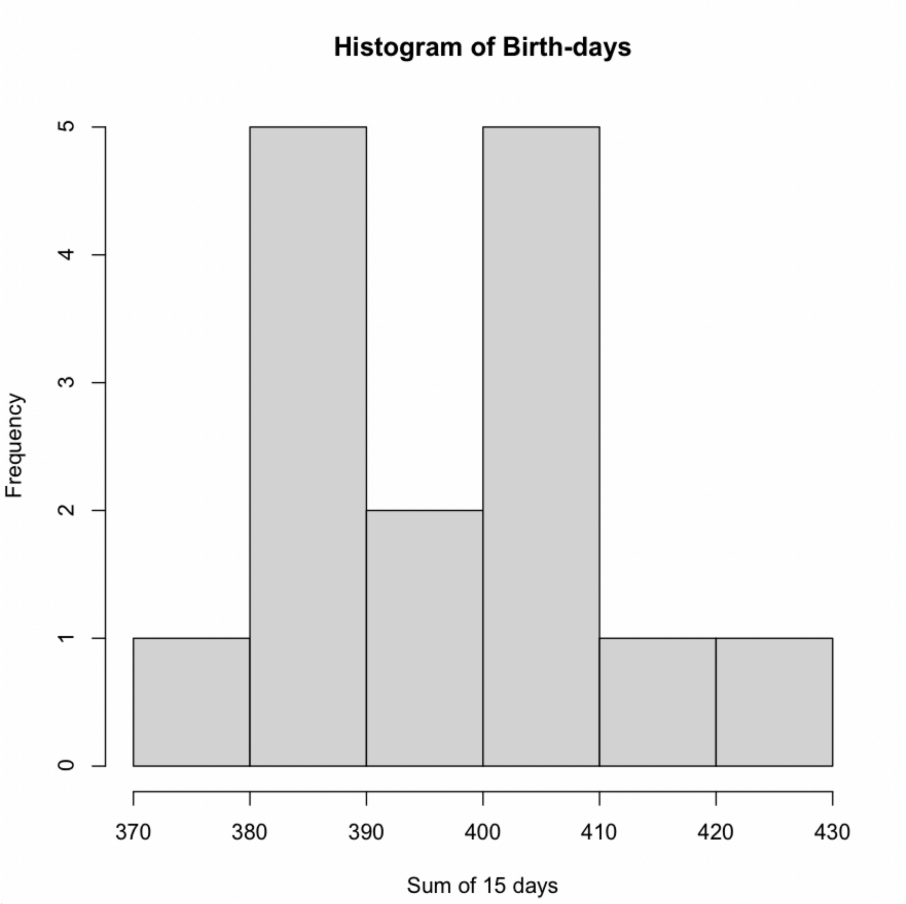
```
setwd("/Users/anshulyadav/Documents/")
ipl<- read.csv("iplbd.csv")

sample_func <- function(){
sample(ipl$Birthday, size=15, replace=TRUE)
}

samples <- replicate(100, sample_func())
sample_sums <- rowSums (samples)

hist(sample_sums, main-"Histogram of Birth-days", xlab="Sum of 15 days")
```

GRAPH-



IV. 100 samples of size 30

Code :

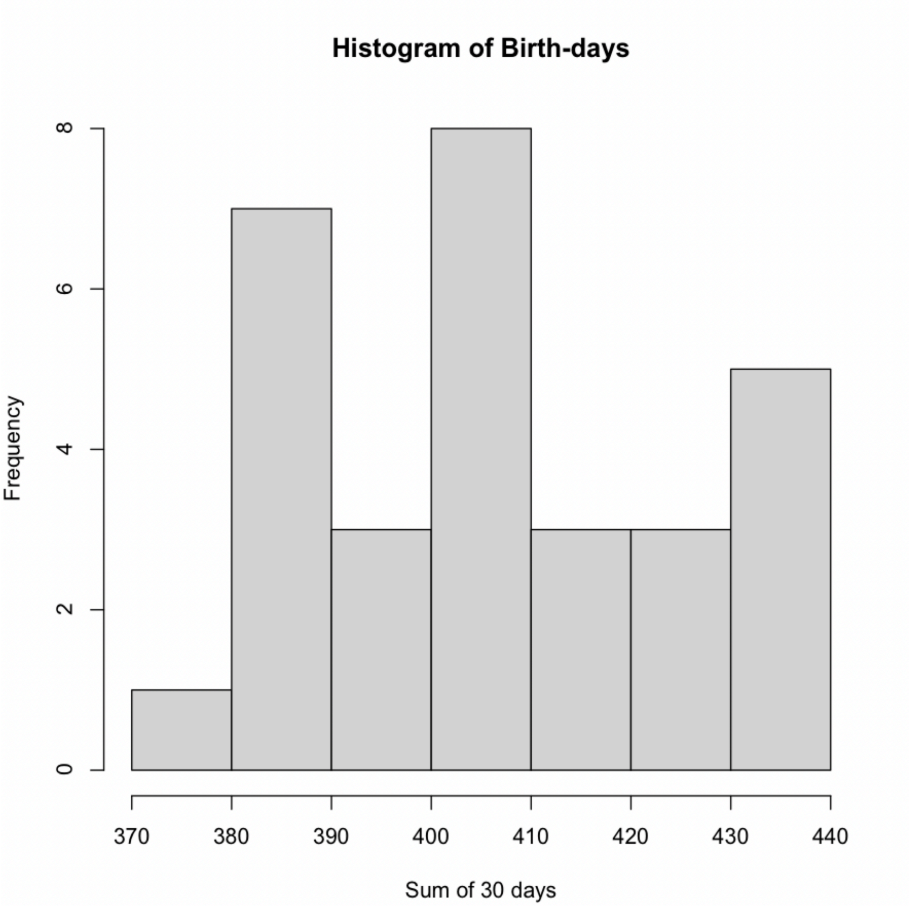
```
setwd("/Users/anshulyadav/Documents/")
ipl<- read.csv("iplbd.csv")

sample_func <- function(){
sample(ipl$Birthday, size=30, replace=TRUE)
}

samples <- replicate(100, sample_func())
sample_sums <- rowSums (samples)

hist(sample_sums, main-"Histogram of Birth-days", xlab="Sum of 30 days")
```

GRAPH-



V. 100 samples of size 50

Code :

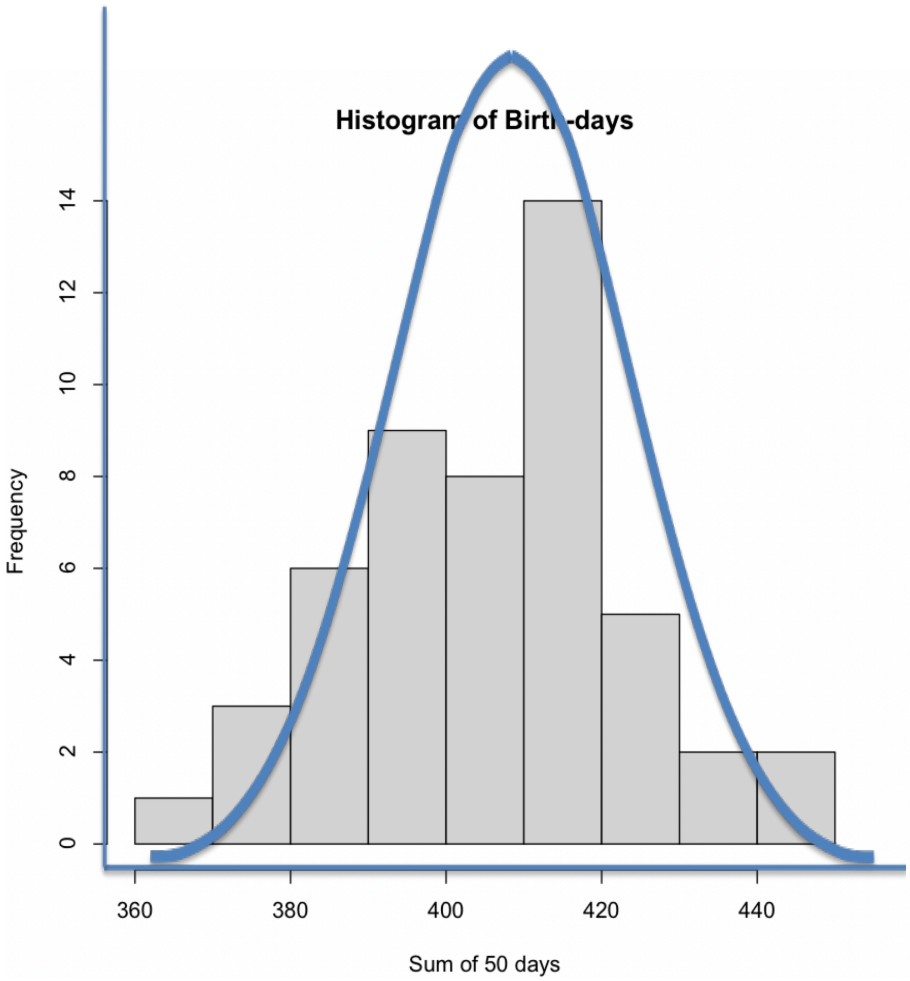
```
setwd("/Users/anshulyadav/Documents/")
ipl<- read.csv("iplbd.csv")

sample_func <- function(){
sample(ipl$Birthday, size=50, replace=TRUE)
}

samples <- replicate(100, sample_func())
sample_sums <- rowSums (samples)

hist(sample_sums, main-"Histogram of Birth-days", xlab="Sum of 50 days")
```

GRAPH-



VI. 50 samples of size 30

Code :

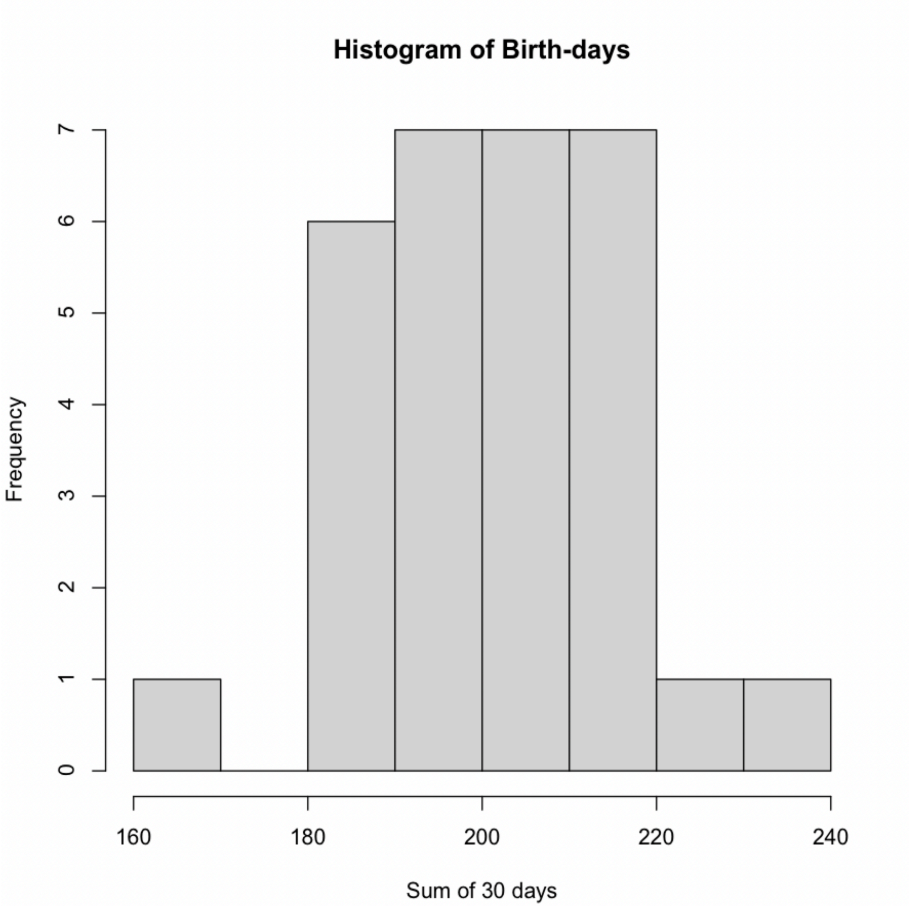
```
setwd("/Users/anshulyadav/Documents/")
ipl<- read.csv("iplbd.csv")

sample_func <- function(){
sample(ipl$Birthday, size=30, replace=TRUE)
}

samples <- replicate(50, sample_func())
sample_sums <- rowSums (samples)

hist(sample_sums, main-"Histogram of Birth-days", xlab="Sum of 30 days")
```

GRAPH-



VII.50 samples of size 50

Code :

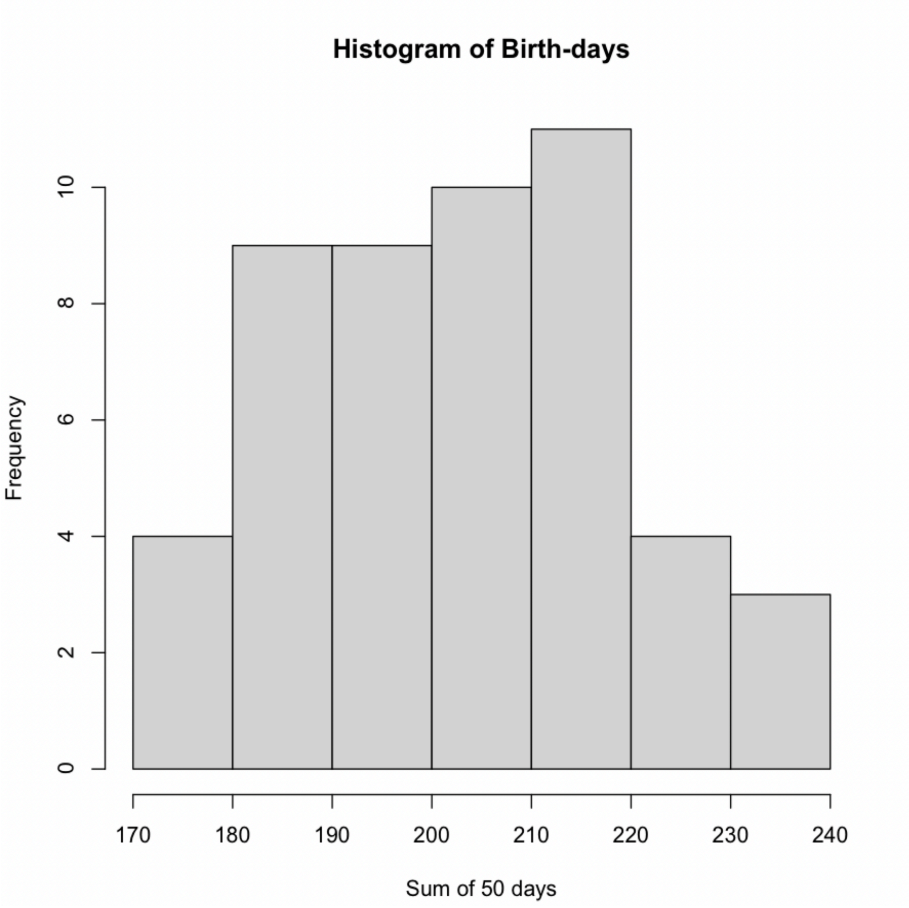
```
setwd(“/Users/anshulyadav/Documents/”)
ipl<- read.csv(“iplbd.csv”)

sample_func <- function(){
sample(ipl$Birthday, size=50, replace=TRUE)
}

samples <- replicate(50, sample_func())
sample_sums <- rowSums (samples)

hist(sample_sums, main-"Histogram of Birth-days“, xlab=“Sum of 50 days”)
```

GRAPH-



VIII.150 samples of size 30

Code :

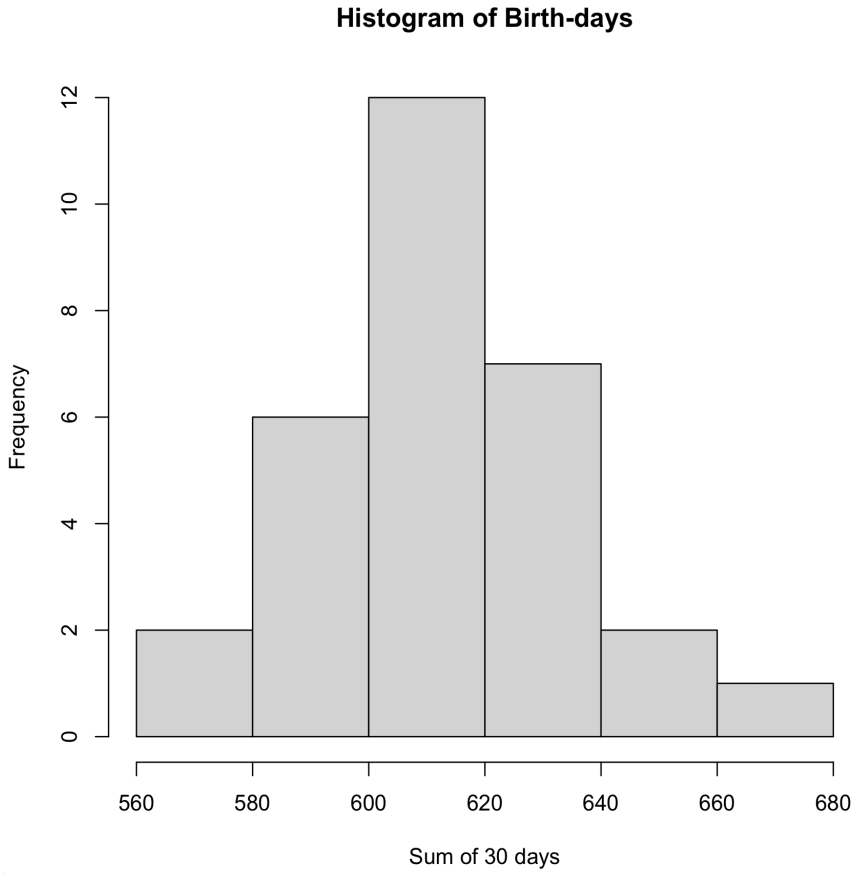
```
setwd("/Users/anshulyadav/Documents/")
ipl<- read.csv("iplbd.csv")

sample_func <- function(){
sample(ipl$Birthday, size=30, replace=TRUE)
}

samples <- replicate(150, sample_func())
sample_sums <- rowSums (samples)

hist(sample_sums, main-"Histogram of Birth-days", xlab="Sum of 30 days")
```

GRAPH-



IX. 150 samples of size 50

Code :

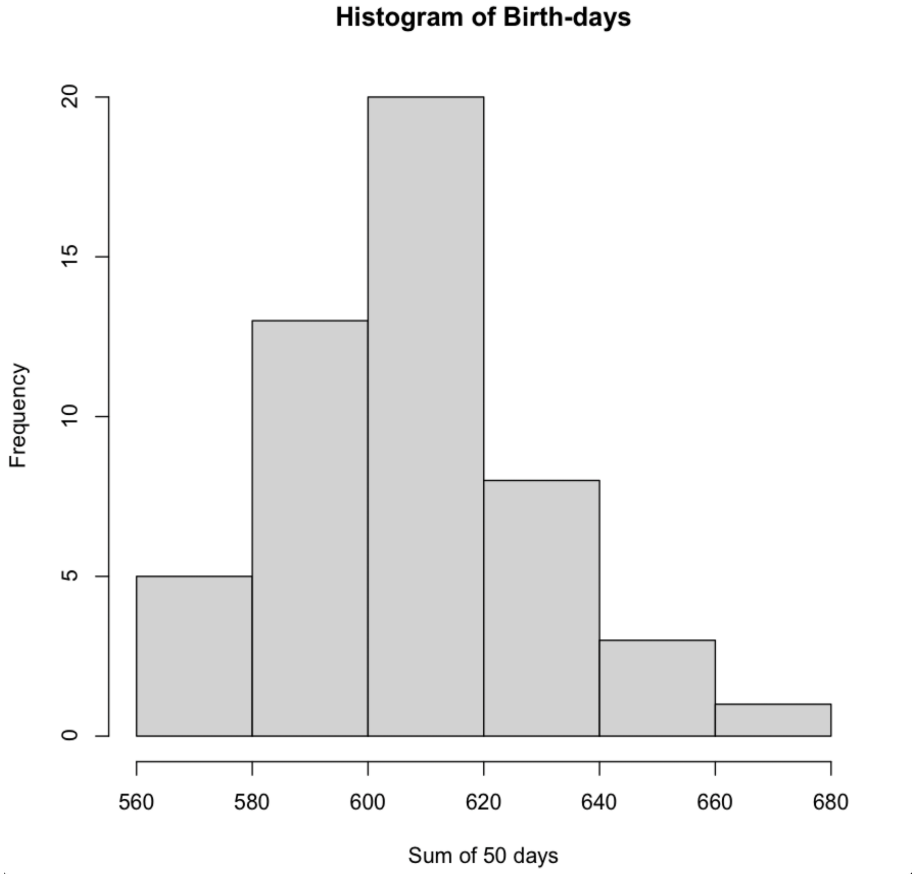
```
setwd("/Users/anshulyadav/Documents/")
ipl<- read.csv("iplbd.csv")

sample_func <- function(){
sample(ipl$Birthday, size=50, replace=TRUE)
}

samples <- replicate(150, sample_func())
sample_sums <- rowSums (samples)

hist(sample_sums, main-"Histogram of Birth-days", xlab="Sum of 50 days")
```

GRAPH-



X. 150 samples of size 100

Code :

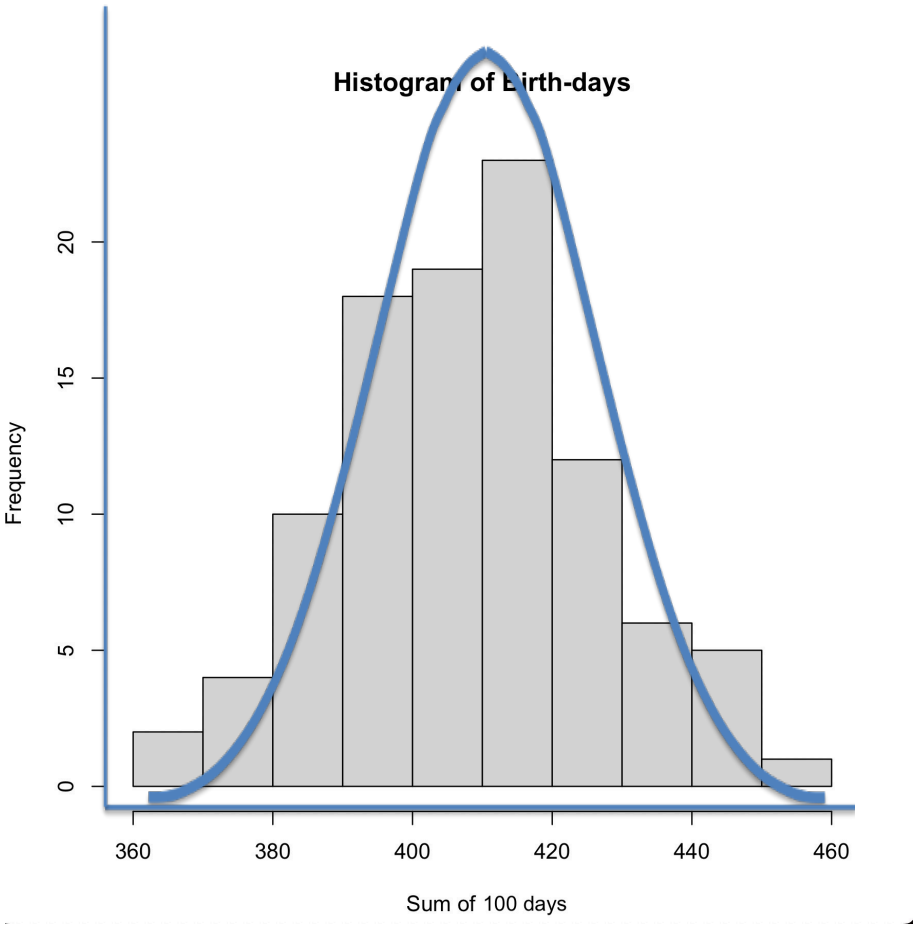
```
setwd("/Users/anshulyadav/Documents/")
ipl<- read.csv("iplbd.csv")

sample_func <- function(){
sample(ipl$Birthday, size=100, replace=TRUE)
}

samples <- replicate(150, sample_func())
sample_sums <- rowSums (samples)

hist(sample_sums, main-"Histogram of Birth-days", xlab="Sum of 100 days")
```

GRAPH-



## **Conclusion** :

A fundamental statistical idea with several practical applications is the central limit theorem. By simulating and analysing a large number of random birthdays and coin tosses, we may investigate how the sum of these tosses and birthdays tends towards a normal distribution, regardless of the underlying distribution of the individual variables.

This displays the CLT's ability to study a large number of random variables' behaviour, which has many uses in statistical analysis and data science.

Using R programming language, we can simulate random birthdays for each day of the week separately and plot their histograms to compare their distributions. From these histograms, we can observe that the distributions of random birthdays on different days of the week are similar, with a roughly bell-shaped curve centred around the mean. This provides further evidence for the validity of the CLT, as we observe how the sum of these random variables tends towards a normal distribution.

Overall, the CLT is a powerful and widely applicable concept in statistics, with practical applications in a wide range of fields. By understanding and utilising this theorem, we can gain a deeper understanding of the behaviour of random variables and improve our ability to analyse and interpret data.

## **Acknowledgement** :

To everyone who helped make this project report a success, we would like to extend our deepest appreciation. First of all, we would like to express our gratitude to Dr. Rishi Asthana, our project manager, for his leadership, support, and direction during the whole project. His insightful observations and constructive criticism helped us mould our work.

We would like to acknowledge the various websites, books, and research papers that have provided valuable data, information, and insights that were crucial to the completion of this project. Without their contribution, this project would not have been possible.

In conclusion, we are grateful to all who have played a part in this project, either directly or indirectly. We hope that the knowledge and findings generated through this study will contribute positively to the field of environmental studies and inspire further research and action towards a more sustainable future for all.

## **References** :


1. https://www.webassign.net/question_assets/idcollabstat2/Chapter7.pdf
2. https://www.analyzemath.com/probabilities/central-limit-theorem-examples-and-solutions.html
3. https://www.probabilitycourse.com/chapter7/7_1_2_central_limit_theorem.php
4. https://www.investopedia.com/terms/c/central_limit_theorem.asp