<u>Practical: - 4 Sampling, Sampling Distribution, Hypothesis Testing</u>

(a) Random sampling *with or without replacement* using **sample()** function. (**Mandatory**)

☐ ***<u>Introduction:-</u>***

The **sample(x, n, replace = FALSE, prob = NULL)** function takes a sample from a vector **x** of size **n**. This sample can be **with** or **without replacement** and the probabilities of selecting each

❖ ***<u>Coin example:-</u>***

sample(x, size = 5)
Output:- `1 2 0 0 3`

Now, let's perform our coin-flipping experiment just once.

coin = c("Heads", "Tails")
sample(coin, size = 1)
Output:- `"Tails"`

And now, let's try it **100** times

sample(coin, size = 100)

```
Error in sample(coin, size = 100) :
  cannot take a sample larger than the population when 'replace = FALSE'
```

Oops, we can't take a sample of size **100** from a vector of size **2**, unless we set the **replace** argument to **TRUE**.
element to the sample can be either **the same for each element** or **a vector** informed by the user.

☐ Tossing **10** coins

sample(0:1, 10, replace = TRUE) Output:- 0 0 1 0 0 1 1 0 0 1

☐ Roll **10** dice

sample(1:6, 10, replace = TRUE)
Output:- 1 4 3 6 1 2 5 5 2 5

☐ Play lottery (**6** random numbers out of **50** *without replacement*)

sample(1:50, 6, replace = FALSE) Output:- 31 15 25 20 22 48

table(sample(coin, size = 100, replace = TRUE))

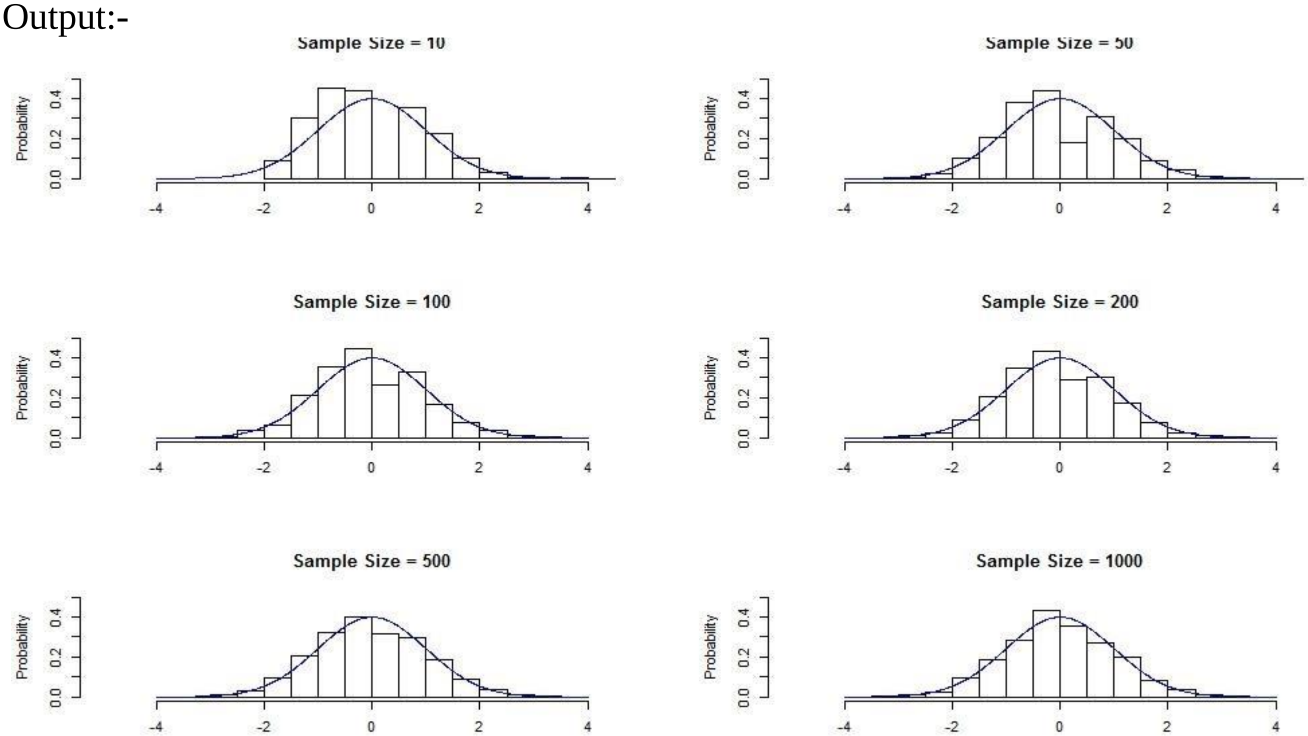
```
Heads Tails
  53    47
```

(b) Generate **n** random samples (take **n = 10, 50, 100, 200, 500, 1000** as an example), create a vector of Sample Means. Draw the **Density Plot** of Sample Means to visualize **Central Limit Theorem**. (**Mandatory**).

Solution:- **p <- 0.05 n <- 6 sims <- 4000**

```
m <- c(10, 50, 100, 200, 500, 1000)
E.of.X <- n*p
V.of.X <- n*p*(1-p)
Z <- matrix(NA, nrow = sims, ncol = length(m)) for (i in
1:sims)
{
        for (j in 1:length(m))
        {
                samp <- rbinom(n = m[j], size = n, prob = p)
sample.mean <- mean(samp)
                Z[i,j] <- (sample.mean - E.of.X) / sqrt(V.of.X/m[j])
        }
}
par(mfrow = c(3,2))

for (j in 1:6)
{ hist(Z[,j], xlim = c(-5, 5), freq = FALSE, ylim = c(0, 0.5), ylab = "Probability", xlab = "",
        main = paste("Sample Size =", m[j]))
        x <- seq(-4, 4, by = 0.01)          y
<- dnorm(x)
        lines(x, y, col = "blue")
}
```

Output:-



(c) Take a sample and carry out Hypothesis Testing for the following cases:

Lower Tail Test of Population Mean with Population Standard Deviation (σ known)

The null hypothesis of the **lower tail test of the population mean** can be expressed as follows:-

$$H_0 : \mu = \mu_0 \quad H_a : \mu < \mu_0$$

where μ_0 is a hypothesized lower bound of the true population mean μ.
Let us define the test statistic z in terms of the sample mean, the sample size and the population standard deviation σ:-

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Problem:-

Suppose the manufacturer claims that the mean lifetime of a light bulb is **less than 10,000** hours. In a **sample** of **30** light bulbs, it was found that they only last **9,900** hours on **average**. Assume the **population standard deviation** is **120** hours. At **0.05** significance level, can we reject the claim by the manufacturer? **Solution:-**

#The null hypothesis is that $\mu \leq 10000$. We begin with computing the test statistic.

```
> xbar = 9900                 # sample mean
> mu0 = 10000                 # hypothesized value
> sigma = 120                 # population standard deviation
> n = 30                      # sample size
> z = (xbar−mu0)/(sigma/sqrt(n))
> z                           # test statistic
```

Output:- -4.564355

#We then compute the **critical value** at **0.05** significance level.

```
> alpha = 0.05
> z.alpha = qnorm(1−alpha)
> −z.alpha                    # critical value
```

Output:- -1.644854

Answer:-

The test statistic **–4.564355** is **less than** the critical value of **–1.644854**. Hence, at 0.05 significance level, we **reject the claim** that mean lifetime of a light bulb is above 10,000 hours. **p-value Method:-**

Instead of using the critical value, we apply the **pnorm()** function to compute the lower tail **p-value** of the test statistic. As it turns out **to be less than the 0.05** significance level, **we reject the null hypothesis** that $\mu \leq 10000$.

```
> pval = pnorm(z)
> pval                        # lower tail p−value
```

Output:- 2.505166e-06

Upper Tail Test of Population Mean with Population Standard Deviation ($\boldsymbol{\sigma}$ known)

The null hypothesis of the **upper tail test of the population mean** can be expressed as follows:-

$$H_0 : \mu = \mu_0 \quad H_a : \mu > \mu_0$$

where μ_0 is a hypothesized upper bound of the true population mean μ.
Let us define the test statistic z in terms of the sample mean, the sample size and the population standard deviation σ:-

$$z = \frac{x - \mu_0}{\sigma/\sqrt{n}}$$

Problem:-

Suppose the food label on a cookie bag states that there is **at least 2** grams of saturated fat in a single cookie. In a **sample** of **35** cookies, it is found that the **mean** amount of saturated fat per cookie is **2.1** grams. Assume that the **population standard deviation** is **0.25** grams. At **0.05** significance level, can we reject the claim on food label?

Solution:-

#The null hypothesis is that $\mu \geq 2$. We begin with computing the test statistic.

```
> xbar = 2.1                # sample mean
> mu0 = 2                   # hypothesized value
> sigma = 0.25              # population standard deviation
> n = 35                    # sample size
> z = (xbar−mu0)/(sigma/sqrt(n))
> z                # test statistic  Output:-  2.366432
```

#We then compute the **critical value** at **0.05** significance level.

```
> alpha = 0.05
> z.alpha = qnorm(1−alpha)
> z.alpha                   # critical value
Output:- 1.644854
```

Answer:-

The test statistic **2.366432** is **greater than** the critical value of **1.644854**. Hence, at 0.05 significance level, we **reject the claim** that there is at most 2 grams of saturated fat in a cookie.

p-value Method:-

Instead of using the critical value, we apply the **pnorm()** function to compute the upper tail **p-value** of the test statistic. As it turns out **to be less than the 0.05** significance level, **we reject the null hypothesis** that $\mu \geq 2$.

```
> pval = pnorm(z, lower.tail=FALSE)
> pval                      # upper tail p−value
```

Output:- 0.008980239

Two-Tailed Test of Population Mean with Population Standard Deviation (σ known)

The null hypothesis of the **two-tailed test of the population mean** can be expressed as follows:-

$$H_0 : \mu = \mu_0 \quad H_a : \mu \neq \mu_0$$

where μ_0 is a hypothesized value of the true population mean μ.
Let us define the test statistic z in terms of the sample mean, the sample size and the population standard deviation σ:-

$$z = \frac{x - \mu_0}{\sigma/\sqrt{n}}$$

Problem:-

Suppose the **mean** weight of King Penguins found in an Antarctic colony last year was **15.4** kg. In a **sample** of **35** penguins same time this year in the same colony, the **mean** penguin weight is **14.6** kg. Assume the **population standard deviation** is **2.5** kg. At **0.05** significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

Solution:-

#The null hypothesis is that μ **= 15.4**. We begin with computing the test statistic.

```
> xbar = 14.6                # sample mean
> mu0 = 15.4                 # hypothesized value
> sigma = 2.5                # population standard deviation
> n = 35                     # sample size
> z = (xbar−mu0)/(sigma/sqrt(n))
> z                # test statistic  Output:-  -1.893146
```

#We then compute the **critical values** at **0.05** significance level.

```
> alpha = 0.05
> z.half.alpha = qnorm(1−alpha/2)  > c(−z.half.alpha,
z.half.alpha)
Output:-  -1.959964  1.959964
```

Answer:-

The test statistic **–1.893146** lies **between** the critical values **–1.959964** and **1.959964**. Hence, at 0.05 significance level, we **do *not* reject the null hypothesis** that the mean penguin weight does not differ from last year.

p-value Method:-

Instead of using the critical value, we apply the **pnorm()** function to compute the two-tailed **p-value** of the test statistic. It doubles the *lower* tail p-value as the sample mean is *less* than the hypothesized value. Since **it turns out to be greater than the 0.05** significance level, we **do *not* reject the null hypothesis** that $\mu = 15.4$

```
> pval = 2 * pnorm(z)        # lower tail
> pval                       # two−tailed p−value
```

Output:- 0.05833852

Lower Tail Test of Population Mean with Population Standard Deviation (σ unknown means s known)

The null hypothesis of the **lower tail test of the population mean** can be expressed as follows:-

$$H_0 : \mu = \mu_0 \quad H_a : \mu < \mu_0$$

where μ_0 is a hypothesized lower bound of the true population mean μ.

Let us define the test statistic t in terms of the sample mean, the sample size and the sample standard deviation *s*:-

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Problem:-

Suppose the manufacturer claims that the **mean** lifetime of a light bulb is **less than 10,000** hours. In a **sample** of **30** light bulbs, it was found that they only last **9,900** hours on **average**. Assume the **sample standard deviation** is **125** hours. At **0.05** significance level, can we reject the claim by the manufacturer?

Solution:-

#The null hypothesis is that $\mu \le 10000$. We begin with computing the test statistic.

```
> xbar = 9900             # sample mean
> mu0 = 10000             # hypothesized value
> s = 125                 # sample standard deviation
> n = 30                  # sample size
> t = (xbar−mu0)/(s/sqrt(n))
> t                 # test statistic  Output:-  -4.38178
```

#We then compute the **critical value** at **0.05** significance level.

```
> alpha = 0.05
> t.alpha = qt(1−alpha, df=n−1)
> −t.alpha                # critical value
Output:-  -1.699127
```

Answer:-

The test statistic **–4.38178** is **less than** the critical value of **–1.699127**. Hence, at 0.05 significance level, **we can reject the claim** that mean lifetime of a light bulb is above 10,000 hours.

p-value Method:-

Instead of using the critical value, we apply the **pt()** function to compute the lower tail **p-value** of the test statistic. As **it turns out to be less than the 0.05** significance level, **we reject the null hypothesis** that $\mu \le 10000$.

```
> pval = pt(t, df=n−1)
> pval                            # lower tail p−value
```

Output:- 7.035026e-05

Upper Tail Test of Population Mean with Population Standard Deviation ($\boldsymbol{\sigma}$ unknown means s known)

The null hypothesis of the **upper tail test of the population mean** can be expressed as follows:-

$$H_0 : \mu = \mu_0$$

$$H_a : \mu > \mu_0$$

where μ_0 is a hypothesized upper bound of the true population mean μ. Let us define the test statistic t in terms of the sample mean, the sample size and the sample standard deviation s:-

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Problem:-

Suppose the food label on a cookie bag states that there is **at least 2** grams of saturated fat in a single cookie. In a **sample** of **35** cookies, it is found that the **mean** amount of saturated fat per cookie is **2.1** grams. Assume that the **sample standard deviation** is **0.3** gram. At **0.05** significance level, can we reject the claim on food label?

Solution:-

#The null hypothesis is that $\mu \geq 2$. We begin with computing the test statistic.

```
> xbar = 2.1                    # sample mean
> mu0 = 2                       # hypothesized value
> s = 0.3                       # sample standard deviation
> n = 35                        # sample size
> t = (xbar−mu0)/(s/sqrt(n))
> t                   # test statistic  Output:-  1.972027
```

#We then compute the **critical value** at **0.05** significance level.

```
> alpha = 0.05
> t.alpha = qt(1−alpha, df=n−1)
> t.alpha             # critical value
Output:-  1.690924
```

Answer:-

The test statistic **1.972027** is **greater than** the critical value of **1.690924**. Hence, at 0.05 significance level, **we can reject the claim** that there is at most 2 grams of saturated fat in a cookie.

p-value Method:-

Instead of using the critical value, we apply the **pt()** function to compute the upper tail **p-value** of the test statistic. As **it turns out to be less than the 0.05** significance level, **we reject the null hypothesis** that $\mu \geq 2$.

```
> pval = pt(t, df=n−1, lower.tail=FALSE)
> pval                                   # upper tail p−value
```

Output:- 0.02839295

Two-Tailed Test of Population Mean with Population Standard Deviation (σ unknown means s known)

The null hypothesis of the **two-tailed test of the population mean** can be expressed as follows:-

$$H_0 : \mu = \mu_0 \quad H_a : \mu \neq \mu_0$$

where μ_0 is a hypothesized value of the true population mean μ.

Let us define the test statistic t in terms of the sample mean, the sample size and the sample standard deviation s:-

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Problem:-

Suppose the **mean** weight of King Penguins found in an Antarctic colony last year was **15.4** kg. In a **sample** of **35** penguins same time this year in the same colony, the **mean** penguin weight is **14.6** kg. Assume the **sample**

standard deviation is **2.5** kg. At **0.05** significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

Solution:-

#The null hypothesis is that μ **= 15.4**. We begin with computing the test statistic.

```
> xbar = 14.6                # sample mean
> mu0 = 15.4                 # hypothesized value
> s = 2.5                    # sample standard deviation
> n = 35                     # sample size
> t = (xbar−mu0)/(s/sqrt(n))
> t                # test statistic  Output:-  -1.893146
```

#We then compute the **critical values** at **0.05** significance level.

```
> alpha = 0.05
> t.half.alpha = qt(1−alpha/2, df=n−1)
> c(−t.half.alpha, t.half.alpha)
```

Output:- -2.032245 2.032245

Answer:-

The test statistic **–1.893146** lies **between** the critical values **–2.032245** and **2.032245**. Hence, at 0.05 significance level, we **do *not* reject the null hypothesis** that the mean penguin weight does not differ from last year.

Alternative Solution:-

Instead of using the critical value, we apply the **pt()** function to compute the twotailed **p-value** of the test statistic. It doubles the *lower* tail p-value as the sample mean is *less* than the hypothesized value. Since **it turns out to be greater than the 0.05** significance level, **we do *not* reject the null hypothesis** that $\mu = 15.4$.

```
> pval = 2 ∗ pt(t, df=n−1)          # lower tail
> pval                               # two−tailed p−value
```

Output: - 0.06687552

Lower Tail Test of Population Variance (Chi-Square Test):-

The null hypothesis of the **lower tail test of the population variance** can be expressed as follows:-

$$H_0: \sigma^2 = \sigma_0^2 \quad H_a: \sigma^2 < \sigma_0^2$$

where σ_0 is a hypothesized upper bound of the true population variance σ^2. Let us define the test statistic chi-square in terms of the sample variance, the sample size and the population variance σ^2:-

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}; \text{ d.f.} = n - 1$$

Problem:-

Highway engineers in Ohio are painting white stripes on a highway. The stripes are supposed to be approximately **10 feet** long. However, because of the machine, the operator, and the motion of the vehicle carrying the equipment, considerable variation occurs among the stripe lengths. Engineers claim that the **variance** of stripes is **not more than 16 inches**. Use the **sample** lengths given here from **12** measured stripes to test the **variance** claim. Assume stripe length is normally distributed. Let α **= 0.05.** The **standard deviation** of the **12** stripes is **5.98544 inches**.

Solution:-

#The null hypothesis is that $\boldsymbol{\sigma^2} \leq \mathbf{16}$. We begin with computing the test statistic.

```
> sigmasq = 16                          # population variance(σ2)
> s = 5.98544                           # sample standard deviation
> ssq = s * s                           # sample variance(s2)
> n = 12                                # sample size
> chisq = ssq*(n−1)/sigmasq
> chisq                   # test statistic Output:-  24.63003
```

#We then compute the **critical value** at **0.05** significance level.

```
> alpha = 0.05
> chisq.alpha = qchisq(1−alpha, df=n−1)
> chisq.alpha             # critical value
```

Output:- 19.67514

Answer:-

The test statistic **24.63003** is **greater than** the critical value of **19.67514**. Hence, at 0.05 significance level, **we can reject the null hypothesis** that there is Engineers claim that the variance of stripes is not more than 16 inches.

Upper Tail Test of Population Variance (Chi-Square Test):-

The null hypothesis of the **upper tail test of the population variance** can be expressed as follows:-

$$H_0: \sigma^2 = \sigma_0^2 \quad H_a: \sigma^2 > \sigma_0^2$$

where σ_0 is a hypothesized upper bound of the true population variance σ^2. Let us define the test statistic chi-square in terms of the sample variance, the sample size and the population variance σ^2:-

$$\boldsymbol{\chi^2 = \frac{(n-1)s^2}{\sigma^2}}; \ \mathbf{d.f. = n - 1}$$

Problem:-

A company produces industrial wiring. One batch of wiring is specified to be **2.16** centimeters (cm) thick. A company inspects the wiring in **7** locations and determines that, on the **average**, the wiring is about **2.16** cm thick. However, the measurements vary. It is unacceptable for the **variance** of the wiring to be **more than 0.04** cm^2. The **standard deviation** of the **7** measurements on this batch of wiring is **0.34** cm. Use $\boldsymbol{\alpha} = \mathbf{0.01}$ to determine whether the **variance** on the sample wiring is too great to meet specifications. Assume wiring thickness is normally distributed.

Solution:-

#The null hypothesis is that $\boldsymbol{\sigma^2} \geq \mathbf{0.04}$. We begin with computing the test statistic.

```
> xbar = 2.16                  # sample mean
> sigmasq = 0.04               # population variance(σ2)
> s = 0.34                     # sample standard deviation
> ssq = s * s                  # sample variance(s2)
> n = 7                        # sample size
> chisq = ssq*(n−1)/sigmasq
> chisq                        # test statistic
Output:- 17.34
#We then compute the critical value at 0.05 significance level.
> alpha = 0.01
> chisq.alpha = qchisq(1−alpha, df=n−1)
> chisq.alpha          # critical value
Output:- 16.81189
```

Answer:-

The test statistic **17.34** is **greater than** the critical value of **16.81189**. Hence, at 0.01 significance level, **we can reject the null hypothesis** that there is the variance on the sample wiring is too great to meet specifications.

Two Tail Test of Population Variance (Chi-Square Test):-

The null hypothesis of the **lower tail test of the population variance** can be expressed as follows:-

$$H_0: \sigma^2 = \sigma_0^2 \quad H_a: \sigma^2 \neq \sigma_0^2$$

where σ_0 is a hypothesized upper bound of the true population variance σ^2. Let us define the test statistic chi-square in terms of the sample variance, the sample size and the population variance σ^2:-

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}; \text{ d.f.} = n - 1$$

Problem:-

A small business has 37 employees. Because of the uncertain demand for its product, the company usually pays overtime on any given week. The company assumed that about **50** total hours of overtime per week is required and that the **variance** on this figure is about **25**. Company officials want to know whether the variance of overtime hours has changed. Given here is a **sample** of **16** weeks of overtime data (in hours per week). Assume hours of overtime are normally distributed. Use these data to test the null hypothesis that the **variance** of overtime data is **25**. Let α **= 0.10.** The **standard deviation** of the **16** weeks of overtime data is **5.30**.

Solution:-

#The null hypothesis is that $\sigma^2 = 25$. We begin with computing the test statistic.

```
> sigmasq = 25                 # population variance(σ2)
> s = 5.30                     # sample standard deviation
> ssq = s * s                  # sample variance(s2)
```

```
> n = 16                                # sample size
> chisq = ssq*(n−1)/sigmasq
> chisq                  # test statistic Output:-  16.854
```
#We then compute the **critical value** at **0.10** significance level.
```
> alpha = 0.10
> chisq.alpha = qchisq(1−alpha/2, df=n−1)
> chisq.alpha                            # critical value
Output:-  24.99579
```

Answer:-

The test statistic **16.854** is **less than** the critical value of **24.99579**. Hence, at 0.10 significance level, **we cannot reject the null hypothesis** that there is the population variance of overtime hours per week is 25.

Practical - 5. Regression and Linear Modeling

(a) Linear regression:- One Independent Variable using **lm()** function; Interpret the output of Model Analysis, Compute Coefficient of Determination(r^2), Interpret results. (**Mandatory**)

 Introduction:-

The general mathematical equation for a linear regression is:- **y = ax + b**

Following is the description of the parameters used:- **y** is the **response** variable.

- **x** is the **predictor** variable.
- **a** and **b** are **constants** which are called the **coefficients**.

lm() Function:-

This function creates the relationship model between the predictor and the response variable.

 Syntax:-

The basic syntax for **lm()** function in linear regression is:- **lm(formula = y ~ x, data)**

Following is the description of the parameters used:-

- **formula** is a symbol presenting the relation between x and y.
- **data** is the vector on which the formula will be applied.

predict() Function:-

The basic syntax for predict() in linear regression is:- **predict(object, newdata)**
Following is the description of the parameters used:-

- **object** is the formula which is already created using the lm() function.
- **newdata** is the vector containing the new value for predictor variable.

Problem:- Develop the equation of the simple regression line to predict sales(y) from advertising(x) expenditures using the given data:-

Advertising(x):-	**12.5**	**3.7**	**21.6**	**60.0**	**37.6**	**6.1**	**16.8**	**41.2**
Sales(y):-	**148**	**55**	**338**	**994**	**541**	**89**	**126**	**379**

Determine the predicted value of **Sales(y)** = ? for **Advertising(x)** = **50**. Compute r^2

Solution:-
x = c(12.5, 3.7, 21.6, 60, 37.6, 6.1, 16.8, 41.2) #create a vector.
y = c(148, 55, 338, 994, 541, 89, 126, 379) #create a vector.
y.lm = lm(y~x)
coeffs = coefficients(y.lm)
coeffs #we get the value of the coefficients of simple regression line Output:-

```
(Intercept)         x
  -46.29181  15.23977
```

newdata = data.frame(x = 50)
predict(y.lm, newdata) #To compute y = ?, when x = 50 is given predict() is used Output:-

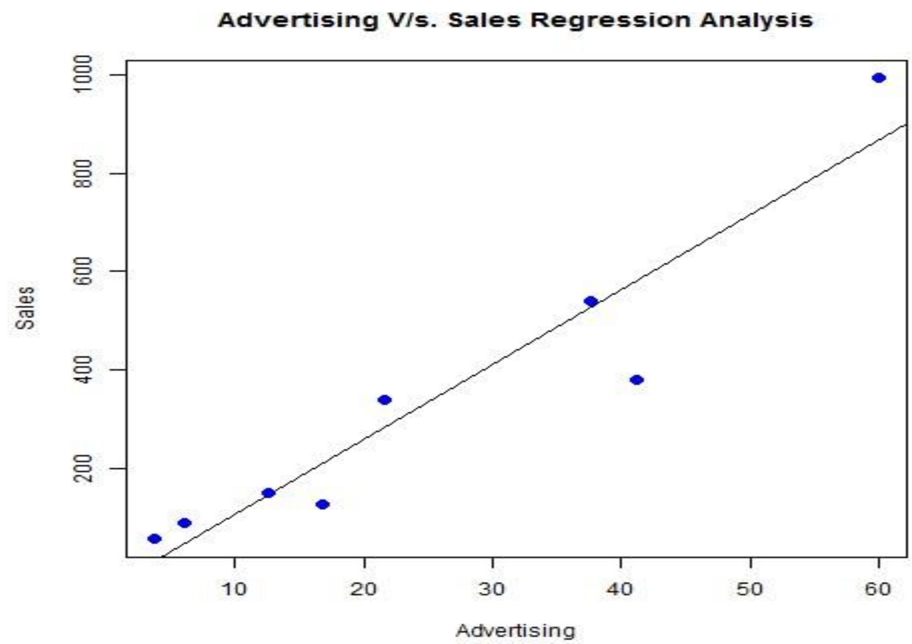
```
1 715.6968
```

print(summary(y.lm)) #It displays the output of Model Analysis.
Output:-

```
Call: lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-202.59  -18.09   28.30   47.46  125.91

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -46.292     64.891  -0.713 0.502402
x             15.240      2.096   7.271 0.000344 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 108.8 on 6 degrees of freedom
Multiple R-squared:  0.8981,  Adjusted R-squared:  0.8811  F-statistic: 52.86 on 1 and 6 DF,  p-value: 0.0003445
```

summary(y.lm)$r.squared #It calculate the value of r^2 separately/directly.
Output:- 0.8980643

png(file = "Linear Regression.png") #create an image of scatter diagram **plot(x, y, col="blue", main="Advertising V/s. Sales Regression Analysis", abline(lm(y~x)), cex = 1.3, pch=16, xlab="Advertising", ylab="Sales")** #use the function of scatterplot
dev.off() #save the file
Output:-



(b) Linear regression:- Multiple Independent Variables using **lm()** function; Interpret the output of Model Analysis. (**Mandatory**)

 Introduction:-

If we choose the parameters α and β_k (k = 1, 2, ..., p) in the multiple linear regression model so as to minimize the sum of squares of the error term ϵ, we will have the so called estimated multiple regression equation. It allows us to compute fitted values of y based on a set of values of x_k (k = 1, 2, ..., p) .

$$\hat{y} = a + \sum_k b_k x_k$$

Following is the description of the parameters used:- **y** is the **response** variable.

- **a, b₁, b₂...bₙ** are the **coefficients**.
- **x₁, x₂, ...xₙ** are the **predictor** variables.

lm() Function:-

This function creates the relationship model between the predictor and the response variable.

Syntax:-

The basic syntax for **lm()** function in multiple regression is:- **lm(formula = y ~ x_1 + x_2 + x_3 ..., data)**
Following is the description of the parameters used:-
 formula is a symbol presenting the relation between x and y. **data** is the vector on which the formula will be applied.

Problem:-
Use a computer to develop the equation of the regression model for the following data. Comment on the regression coefficients. Determine the predicted value of **y** for **x_1 = 33**, **x_2 = 29** and **x_3 = 13**.

y	x_1	x_2	x_3
114	21	6	5
94	43	25	8
87	56	42	25
98	19	27	9
101	29	20	12
85	34	45	21
94	40	33	14
107	32	14	11
119	16	4	7
93	18	31	16
108	27	12	10
117	31	3	8

Solution:- **y = c(114, 94, 87, 98, 101, 85, 94, 107, 119, 93, 108, 117)** #create a vector. **x_1 = c(21, 43, 56, 19, 29, 34, 40, 32, 16, 18, 27, 31)** #create a vector. **x_2 = c(6, 25, 42, 27, 20, 45, 33, 14, 4, 31, 12, 3)** #create a vector. **x_3 = c(5, 8, 25, 9, 12, 21, 14, 11, 7, 16, 10, 8)** #create a vector. **y.lm = lm(y ~ x_1 + x_2 + x_3) coeffs = coefficients(y.lm)**
coeffs #we get the value of the coefficients of multiple regression model Output:-

```
(Intercept)          x1          x2          x3
118.55951024  -0.07940245  -0.88428115   0.37690982
```

newdata = data.frame(x_1 = 33, x_2 = 29, x_3 = 13)
predict(y.lm, newdata) #To compute y = ?, when x_1 = 33, x_2 = 29, x_3 = 13 is given predict() is used Output:-

```
       1
 95.1949
```

print(summary(y.lm)) #It displays the output of Model Analysis. Output:-

```
Call: lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7481 -1.6934  0.5343  1.1214  2.6097

Coefficients:
            Estimate Std. Error t value Pr(>|t|)     (Intercept) 118.55951    1.85798  63.811 4.05e-12 ***
x1          -0.07940    0.06848  -1.159    0.280     x2          -0.88428    0.08631 -10.245 7.08e-06 *** x3
0.37691    0.21973   1.715    0.125
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 2.134 on 8 degrees of freedom
Multiple R-squared: 0.975, Adjusted R-squared: 0.9656 F-statistic: 103.8 on 3 and 8 DF, p-value: 9.582e-07

summary(y.lm)$r.squared **#**It calculate the value of **r^2** separately/directly. Output:-
0.9749568