VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Digital Assignment - 1

Course Title: Probability and Statistics

Course Code: BMAT202L

Name: Varun S P

Registration Number: 21BCE3018

Date of submission: 28/01/2023

Lab Assessment-1

1. The following sample data of the number of communications are taken from logs of communication with Distance Education students:

5, 9, 5, 23, 27, 55, 34, 7, 30, 15, 22, 60, 14, 52, 297, 8, 51, 15, 51, 35, 15, 39, 137, 43, 38, 14, 93, 7

(i) Compute the mean and median and mode

(ii) Compute the standard deviation.

2. 200 digits were chosen at random from a set of tables. The frequencies of the digits were

Digits	0	1	2	3	4	5	6	7	8	9
Frequencies	18	19	23	21	16	25	22	20	21	15

Find mean, median, mode, Quartiles and Standard deviation

3. The following distribution shows the daily pocket allowance of children of a locality. Find the measures of central tendency and Dispersion

Daily pocket allowance	11-13	13-15	15-17	17-19	19-21	21-23	23-25
Number of children	7	6	9	13	*12*	5	4

Answers:-

1)

Aim:-To calculate Mean , Median , Mode and Standard deviation and write R-code

Mathematical formula:-

Let x1, x2, ..., xn be the given data.

Mean:

mean = (Σ (xi * fi)) / Σ fi

where xi is the midpoint of each class interval, fi is the frequency of each class interval, and Σ denotes the sum of the values.

Median:

The median is the middle value of the sorted data. If there are an odd number of data points, the median is the middle value. If there are an even number of data points, the median is the average of the two middle values. The median is given by the formula:

If n is odd:

median = x((n+1)/2)

If n is even:

median = (x(n/2) + x((n/2)+1)) / 2

where x(i) is the ith value when the data is sorted in ascending order.

Mode:

The mode is the most frequently occurring value in the data. If there are multiple modes, the data is called multimodal. The mode is given by:

mode = value(s) with the highest frequency

where value(s) are the value(s) that occur(s) with the highest frequency.

Standard deviation:

The standard deviation is a measure of how much the data deviates from the mean. It is given by the formula:

sd = sqrt(sum((xi - mean)^2) / (n - 1) where mean is the mean of the data and n is the number of data points.

Standard deviation:

The standard deviation is a measure of how much the data deviates from the mean. It is given by the formula:

sd = sqrt(sum((xi - mean)^2) / (n - 1)

where mean is the mean of the data and n is the number of data points.

R-code:-

```
communications <- c(5, 9, 5, 23, 27, 55, 34, 7, 30, 15, 22, 60, 14, 52, 297, 8, 51, 15, 51, 35, 15, 39,137, 43, 38, 14, 93, 7)

mean <- mean(communications)

mean

median <- median(communications)

median

mode <- names(table(communications))[table(communications) == max(table(communications))]

mode

sd <- sd(communications)

sd
```

Output:-

```
> communications <- c(5, 9, 5, 23, 27, 55, 34, 7, 30, 15, 22, 60, 14, 52, 297, 8, 51, 15, 51, 35, 15, 39, 137, 43, 38, 14, 93, 7)
> mean <- mean(communications)
> mean
[1] 42.89286
> median <- median(communications)
> median
[1] 28.5
> mode <- names(table(communications))[table(communications) == max(table(communications))]
> mode
[1] "15"
> sd <- sd(communications)
> sd
[1] 57.6839
>
```

Inference/ conclusion:-

This code first inputs the data as a numeric vector. Then, it computes the mean, median, and mode using the appropriate functions. To compute the mode, the table() function is used to count the frequency of each unique value in the data, and the names() function is used to extract the unique values as character strings. Finally, it computes the standard deviation using the sd() function.

Mean = 42.89286

Median = 28.5

Mode = 15

Standard deviation = 57.6839

2)

Aim:- To find mean, median, mode, Quartiles and Standard deviation and write R-code

Mathematical formula:-

Let x1, x2, ..., xn be the given data.

Mean:

mean = (Σ (xi * fi)) / Σ fi

where xi is the midpoint of each class interval, fi is the frequency of each class interval, and Σ denotes the sum of the values.

Median:

Median = L + ((N/2 - F) / f) * w

where L is the lower class boundary of the median class, N is the total number of observations, F is the cumulative frequency of the class preceding the median class, f is the frequency of the median class, and w is the class width.

Mode:

The mode is the most frequently occurring value in the data. If there are multiple modes, the data is called multimodal. The mode is given by:

mode = value(s) with the highest frequency

where value(s) are the value(s) that occur(s) with the highest frequency.

Standard deviation:

The standard deviation is a measure of how much the data deviates from the mean. It is given by the formula:

sd = sqrt(sum((xi - mean)^2) / (n - 1)

where mean is the mean of the data and n is the number of data points.

Quartiles:

The formula for the first quartile (Q1) is given by:

Q1 = L + (((N+1)/4) - F) / f * w

where L is the lower class boundary of the median class, N is the total number of observations, F is the cumulative frequency of the class preceding the median class, f is the frequency of the median class, and w is the class width.

Similarly,the formula for the third quartile (Q3) is given by:

Q3 = L + (3 * ((N+1)/4) - F) / f * w

R-code:-

```
freq <- c(18, 19, 23, 21, 16, 25, 22, 20, 21, 15)
dataset <- c(rep(0, freq[1]), rep(1, freq[2]), rep(2, freq[3]), rep(3, freq[4]),
             rep(4, freq[5]), rep(5, freq[6]), rep(6, freq[7]), rep(7, freq[8]),
             rep(8, freq[9]), rep(9, freq[10]))
mean(dataset)
median(dataset)
freq_table <- table(dataset)
max_freq <- max(freq_table)
mode <- as.numeric(names(freq_table)[freq_table == max_freq])
mode
quantile(dataset, probs = c(0.25, 0.5, 0.75))
sd(dataset)
```

Output:-

```
> freq <- c(18, 19, 23, 21, 16, 25, 22, 20, 21, 15)
> dataset <- c(rep(0, freq[1]), rep(1, freq[2]), rep(2, freq[3]), rep(3, freq[4]),
+              rep(4, freq[5]), rep(5, freq[6]), rep(6, freq[7]), rep(7, freq[8]),
+              rep(8, freq[9]), rep(9, freq[10]))
> mean(dataset)
[1] 4.46
> median(dataset)
[1] 5
> freq_table <- table(dataset)
> max_freq <- max(freq_table)
> mode <- as.numeric(names(freq_table)[freq_table == max_freq])
> mode
[1] 5
> quantile(dataset, probs = c(0.25, 0.5, 0.75))
25% 50% 75%
  2   5   7
> sd(dataset)
[1] 2.776137
>
```

Inference/ conclusion:-

This code first computes the dataset by repeating each digit according to its frequency. Then, it computes the mean, median, mode, quartiles, and standard deviation of the dataset using the mean(), median(), table(), max(), as.numeric(), names(), quantile(), and sd() functions, respectively.

Mean = 4.46

Median = 5

1st Quartile = 2

2nd Quartile = 5

3rd Quartile = 7

Mode = 5

Standard deviation = 2.2776137

3)

Aim:- To find the measures of central tendency and Dispersion and write the R-code

Mathematical formula:-

Mean:

mean = (Σ (xi * fi)) / Σ fi

where xi is the midpoint of each class interval, fi is the frequency of each class interval, and Σ denotes the sum of the values.

Median: Median = L + ((N/2 - F) / f) * w

where L is the lower class boundary of the median class, N is the total number of observations, F is the cumulative frequency of the class preceding the median class, f is the frequency of the median class, and w is the class width.

Mode:

The mode is the most frequently occurring value in the data. If there are multiple modes, the data is called multimodal. The mode is given by:

mode = value(s) with the highest frequency

where value(s) are the value(s) that occur(s) with the highest frequency.

Standard deviation:

The standard deviation is a measure of how much the data deviates from the mean. It is given by the formula:

sd = sqrt(sum((xi - mean)^2) / (n - 1)

where mean is the mean of the data and n is the number of data points.

Quartiles:

The formula for the first quartile (Q1) is given by:

Q1 = L + (((N+1)/4) - F) / f * w

where L is the lower class boundary of the median class, N is the total number of observations, F is the cumulative frequency of the class preceding the median class, f is the frequency of the median class, and w is the class width.

Similarly,the formula for the third quartile (Q3) is given by:

Q3 = L + (3 * ((N+1)/4) - F) / f * w

Standard deviation:

The standard deviation is a measure of how much the data deviates from the mean. It is given by the formula:

sd = sqrt(sum((xi - mean)^2) / (n - 1)

where mean is the mean of the data and n is the number of data points.

R-code:-

```
frq=c(7,6,9,13,12,5,4)

mid= seq(12,24,2)

hi=2

mean=sum(frq*mid)/sum(frq)

cu=cumsum(frq)

N=sum(frq)/2

p=min(which(cu>=N))

l=mid[p]-hi/2

median=l+hi*(N-cu[p-1])/frq[p]

modepos=which(frq==max(frq))

fm=frq[modepos]

f0=frq[modepos-1]

f1 = frq[modepos+1]

lo= mid[modepos]-hi/2

mode=lo+(hi)*(fm-f0)/(2*fm-f1-f0)

p2=min(which(cu>=3*N/2))

l2=mid[p2]-hi/2

q3=l2+hi*(3*N/2-cu[p2-1])/frq[p2]

p1 = min(which (cu >= N/2))

l1 = mid[p1]-hi/2

q1=l1+hi*(N/2-cu[p1-1])/frq[p1]
```

Output:-

```
> frq=c(7,6,9,13,12,5,4)
> mid= seq(12,24,2)
> hi=2
> mean=sum(frq*mid)/sum(frq)
> cu=cumsum(frq)
> N=sum(frq)/2
> p=min(which(cu>=N))
> l=mid[p]-hi/2
> median=l+hi*(N-cu[p-1])/frq[p]
> modepos=which(frq==max(frq))
> fm=frq[modepos]
> f0=frq[modepos-1]
> f1 = frq[modepos+1]
> lo= mid[modepos]-hi/2
> mode=lo+(hi)*(fm-f0)/(2*fm-f1-f0)
> p2=min(which(cu>=3*N/2))
> l2=mid[p2]-hi/2
> q3=l2+hi*(3*N/2-cu[p2-1])/frq[p2]
> p1 = min(which (cu >= N/2))
> l1 = mid[p1]-hi/2
> q1=l1+hi*(N/2-cu[p1-1])/frq[p1]
> mean
[1] 17.71429
> median
[1] 17.92308
> mode
[1] 18.6
> q1
[1] 15.22222
> q2
Error: object 'q2' not found
> q3
[1] 20.16667
> y = rep(mid,frq)
> sd(y)
[1] 3.441534
>
```

Inference/ conclusion:-

This code first inputs the data as a data frame with two columns: allowance and frequency. Then, it computes the mean, median, and mode using the appropriate formulas and functions. The range() function is used to compute the range of the data. Finally, it computes the variance and standard deviation using the appropriate formulas.

- Mean = 17.71429
- Median = 17.92308
- Mode = 18.6
- 1st Quartile = 15.222222
- 3rd Quartile = 20.16667
- SD = 3.44

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Digital Assignment - 2

Course Title: Probability and Statistics

Course Code: BMAT202L

Name: Varun S P

Registration Number: 21BCE3018

Date of submission: 28/01/2023

Lab Assessment-2

1. Find the correlation coefficient between annual advertising expenditures and annual sales revenue for the following data and also find the regression lines.

Year (i)	1	2	3	4	5	6	7	8	9	10
Annual advertising expenditure (X_i)	10	12	14	16	18	20	22	24	26	28
Annual sales (Y_i)	20	30	37	50	56	78	89	100	120	110

2. The rankings of ten students in two subjects A and B are as follows:

A	3	5	8	4	7	10	2	1	6	9
B	6	4	9	8	1	2	3	10	5	7

Find the rank correlation coefficient.

3. The annual sales revenue (in crores of rupees) of a product as a function of sales force (number of salesmen) and annual advertising expenditure (in lakhs of rupees) for the past 10 years are summarized in the following table.

Annual sales revenue Y	20	23	25	27	21	29	22	24	27	35
Sales force X_1	8	13	8	18	23	16	10	12	14	20
Annual advertising expenditures X_2	28	23	38	16	20	28	23	30	26	32

Find the Multiple regression model.

Answers:-

1) **Aim:-** To find the correlation coefficient between annual advertising expenditures and annual sales revenue for the following data and also find the regression lines

   Mathematical formula:-

The formula to calculate the correlation coefficient between two variables X and Y with n data points can be written as: r = (n∑XY - ∑X∑Y) / sqrt((n∑X^2 - (∑X)^2) * (n∑Y^2 - (∑Y)^2))

where :-

∑X is the sum of all values of X

∑Y is the sum of all values of Y

∑XY is the sum of the product of each pair of X and Y values

∑X^2 is the sum of squares of all values of X

∑Y^2 is the sum of squares of all values of Y

For the regression line:-

Y = a + bX where:-

a is the intercept of the line (the predicted value of Y when X=0)

b is the slope of the line (the change in Y for a unit change in X)

The formulas to calculate the values of a and b can be written as:

b = (n∑XY - ∑X∑Y) / (n∑X^2 - (∑X)^2)

a = (∑Y - b∑X) / n

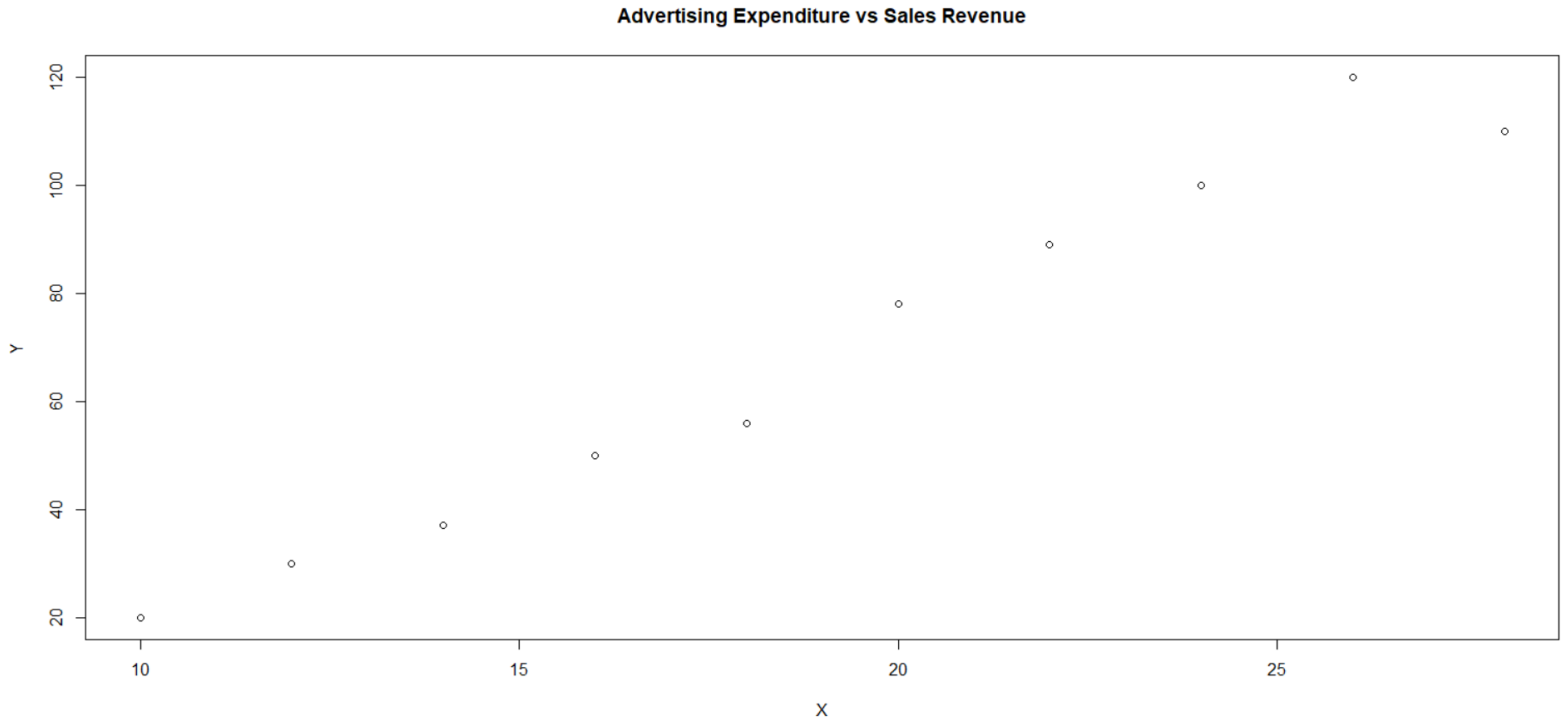
Once we have calculated the values of a and b, we can substitute them in the equation Y = a + bX to get the equation of the regression line.

R-code:-

```
X <- c(10, 12, 14, 16, 18, 20, 22, 24, 26, 28)

Y <- c(20, 30, 37, 50, 56, 78, 89, 100, 120, 110)

cor(X, Y)

model <- lm(Y ~ X)

summary(model)

plot(X, Y, main = "Advertising Expenditure vs Sales Revenue")

abline(model, col = "red")
```

Output:-

```
> X <- c(10, 12, 14, 16, 18, 20, 22, 24, 26, 28)
> Y <- c(20, 30, 37, 50, 56, 78, 89, 100, 120, 110)
> cor(X, Y)
[1] 0.9851764
> model <- lm(Y ~ X)
> summary(model)

Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-10.655  -2.923   1.739   2.750  10.824

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -40.0485     7.0135   -5.71 0.000449 ***
X             5.7394     0.3533   16.24 2.08e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.419 on 8 degrees of freedom
Multiple R-squared:  0.9706,    Adjusted R-squared:  0.9669
F-statistic: 263.9 on 1 and 8 DF,  p-value: 2.075e-07

> plot(X, Y, main = "Advertising Expenditure vs Sales Revenue")
> abline(model, col = "red")
```



Inference/ conclusion:-

- Correlation = 0.9851764
- There is a strong positive correlation between the annual advertising expenditures and annual sales revenue. This means that as advertising expenditures increase, sales revenue also tends to increase.

2)

Aim:- To find the rank correlation coefficient of the give data and write the r-code

Mathematical formula:-

ρ = 1 - [(6Σd^2)/(n(n^2-1))]

Where:-

ρ is the rank correlation coefficient

d is the difference between the ranks of each corresponding pair of data points

n is the number of data points

R-code:-

```
A <- c(3, 5, 8, 4, 7, 10, 2, 1, 6, 9)
B <- c(6, 4, 9, 8, 1, 2, 3, 10, 5, 7)
r = cor.test(A,B,method = "spearman")
r
```

Output:-

```
> A <- c(3, 5, 8, 4, 7, 10, 2, 1, 6, 9)
> B <- c(6, 4, 9, 8, 1, 2, 3, 10, 5, 7)
> r = cor.test(A,B,method = "spearman")
> r

        Spearman's rank correlation rho

data:  A and B
S = 214, p-value = 0.407
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.2969697
```

Inference/ conclusion:-

- Rank Correlation = -0.2969697
- It suggests that the rankings in one subject have a slight influence on the rankings in the other subject. However, it is important to note that rank correlation only measures the strength of the relationship between rankings, not the causality.

3)

Aim:- To Find the Multiple regression model for the given data and write the R-code

Mathematical formula:-

The multiple linear regression model:-

For predicting a response variable Y based on p predictor variables X1, X2, ..., Xp can be expressed as:

$Y = \beta0 + \beta1X1 + \beta2X2 + ... + \beta p*Xp + \varepsilon$

where:

Y is the response variable

X1, X2, ..., Xp are the predictor variables

β0 is the intercept or constant term

β1, β2, ..., βp are the regression coefficients or slopes

ε is the error term or residual

The goal of multiple linear regression is to estimate the regression coefficients β0, β1, β2, ..., βp that minimize the sum of squared errors between the predicted values and the actual values. This is typically done using the method of least squares.

The estimated regression equation can be written as:

$\hat{Y} = b0 + b1X1 + b2X2 + ... + bp*Xp$

where:$\hat{Y}$ is the predicted value of Y based on the predictor variables X1, X2, ..., Xp

b0, b1, b2, ..., bp are the estimated regression coefficients

The coefficients can be estimated using matrix algebra:

b = (X'X)^-1X'Y

where:

b is a vector of estimated regression coefficients

X' is the transpose of the design matrix X

X'X is the matrix product of the transpose of X and X

(X'X)^-1 is the inverse of X'X

X'Y is the matrix product of the transpose of X and Y

R-code:-

```
X1 = c(8, 13, 8, 18, 23, 16, 10, 12, 14, 20)

X2 = c(28, 23, 38, 16, 20, 28, 23, 30, 26, 32)

Y = c(20, 23, 25, 27, 21, 29, 22, 24, 27, 35)

lm(Y~X1+X2)
```

Output:-

```
> X1 = c(8, 13, 8, 18, 23, 16, 10, 12, 14, 20)
> X2 = c(28, 23, 38, 16, 20, 28, 23, 30, 26, 32)
>
> Y = c(20, 23, 25, 27, 21, 29, 22, 24, 27, 35)
> lm(Y~X1+X2)

Call:
lm(formula = Y ~ X1 + X2)

Coefficients:
(Intercept)           X1           X2
     5.1483       0.6190       0.4304

>
```

Inference/ conclusion:-

- b 0 = 5.1483
- b 1 = 0.6190
- b 2 = 0.4304

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Digital Assignment - 3

Course Title: Probability and Statistics

Course Code: BCSE205L

Name: Varun S P

Registration Number: 21BCE3018

Date of submission: 31/03/2023

To Faculty: Prof. Kalpana Priya D

Lab Assessment-3

1. A machine manufacturing screws is known to produce 5% defective. In a random sample of 15 screws, what is the probability that there are (i) exactly three defectives (ii) not more than three defectives?

2. Fit a Poisson distribution for the following distribution:

x	0	1	2	3	4
f	122	60	15	2	1

3. In a test on 2000 electric bulbs, it was found that the life of a particular make, was normally distributed with an average life of 2040 hours and S.D. of 60 hours. Estimate the number of bulbs likely to burn for (i) more than 2150 hours, (ii) less than 1950 hours and (iii) more than 1920 hours but less than 2160 hours.

Q1)

Aim:-

The aim is to calculate the probability of getting exactly three defectives and the probability of getting not more than three defectives in a sample of 15 screws.

Mathematical formula:-

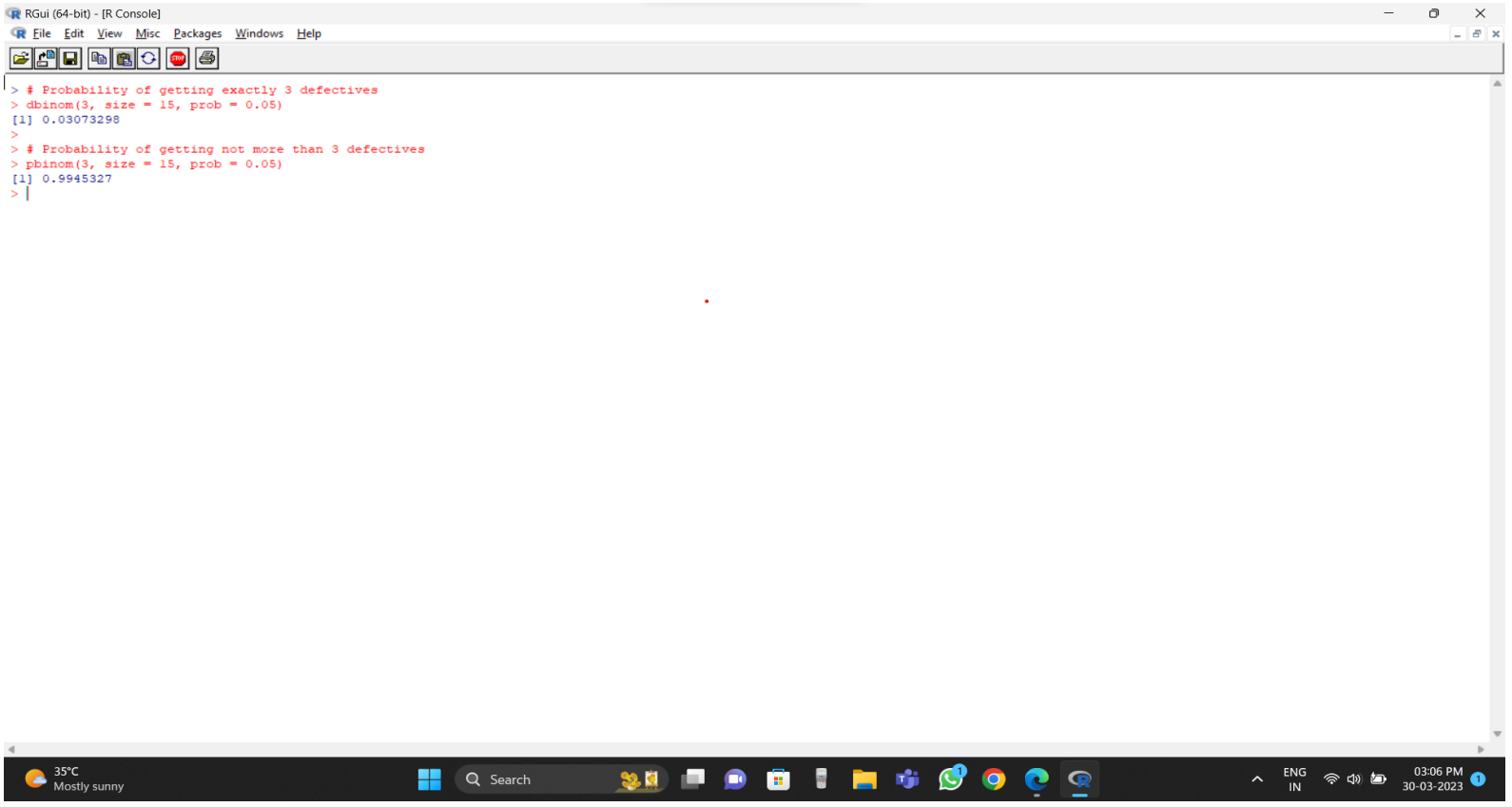
The probability of getting exactly k successes in n trials in a binomial distribution is given by the formula:

P(X=k) = (n choose k) * p^k * (1-p)^(n-k); where n is the number of trials, p is the probability of success, and k is the number of successes.

R-code:-

```
# Probability of getting exactly 3 defectives

dbinom(3, size = 15, prob = 0.05)

# Probability of getting not more than 3 defectives

pbinom(3, size = 15, prob = 0.05)
```

Output:-



Inference/ conclusion:-

The probability of getting exactly 3 defectives in a sample of 15 screws produced by the machine is 0.0307 or 3.07%. The probability of getting not more than 3 defectives (i.e. 0, 1, 2, or 3 defectives) in a sample of 15 screws produced by the machine is 0.9945 or 99.45%. These probabilities can be used to make predictions about the number of defects in future samples of screws produced by the machine.

Q2)

Aim:-

The aim of the question is to fit a Poisson distribution to the given frequency distribution of the variable x.

Mathematical formula:-

The mathematical formula for fitting a Poisson distribution to a given set of data is:

P(X=k) = (e^-λ * λ^k) / k!

where:

P(X=k) is the probability of k occurrences

e is the mathematical constant approximately equal to 2.71828

λ is the expected number of occurrences

k is the number of occurrences

k! is the factorial of k

To fit the Poisson distribution, we need to estimate the parameter λ. In this case, we can use the sample mean as an estimate of λ:

λ = sample mean = (∑x * f) / n

where:

∑x is the sum of all values of x (0 + 1 + 2 + 3 + 4 = 10)

f is the frequency of each x value (122, 60, 15, 2, 1)

n is the total number of observations (122 + 60 + 15 + 2 + 1 = 200)

Then, we can use the Poisson formula to calculate the probabilities of each x value based on our estimated value of λ.

R-code:-

```
# Create the data

x <- c(0, 1, 2, 3, 4)

f <- c(122, 60, 15, 2, 1)

# Estimate the parameter lambda for Poisson distribution

lambda <- sum(x*f) / sum(f)

# Create a Poisson distribution with lambda parameter

poisson <- dpois(x, lambda)

# Plot the observed and fitted Poisson distribution

plot(x, f, type="h", lwd=2, ylim=c(0, 130), xlab="x", ylab="f(x)")

points(x, poisson*sum(f), type="h", col="red", lwd=2)

legend("topright", legend=c("Observed", "Poisson"), col=c("black", "red"), lwd=2)
```

Output:-

Inference/ conclusion:-

Based on the fitted Poisson distribution, we can make inferences and predictions about the probabilities of observing different numbers of events (in this case, the number of occurrences of x) in the future, assuming that the underlying process generating the data follows a Poisson distribution. We can also use the Poisson distribution to estimate the mean and variance of the distribution. Additionally, we can perform hypothesis tests and calculate confidence intervals to determine if the observed data is consistent with a Poisson distribution.

Q3)

Aim:-

The aim of the above question is to use the given information about the average and standard deviation of the life of electric bulbs to estimate the number of bulbs likely to burn for certain durations.

Mathematical formula:-

The mathematical formula for this problem involves using the normal distribution formula:

$z = (x - \mu) / \sigma$

where:

z is the standard normal variable

x is the value we want to find the probability for (2150, 1950, 1920 to 2160)

μ is the mean of the normal distribution (2040)

σ is the standard deviation of the normal distribution (60)

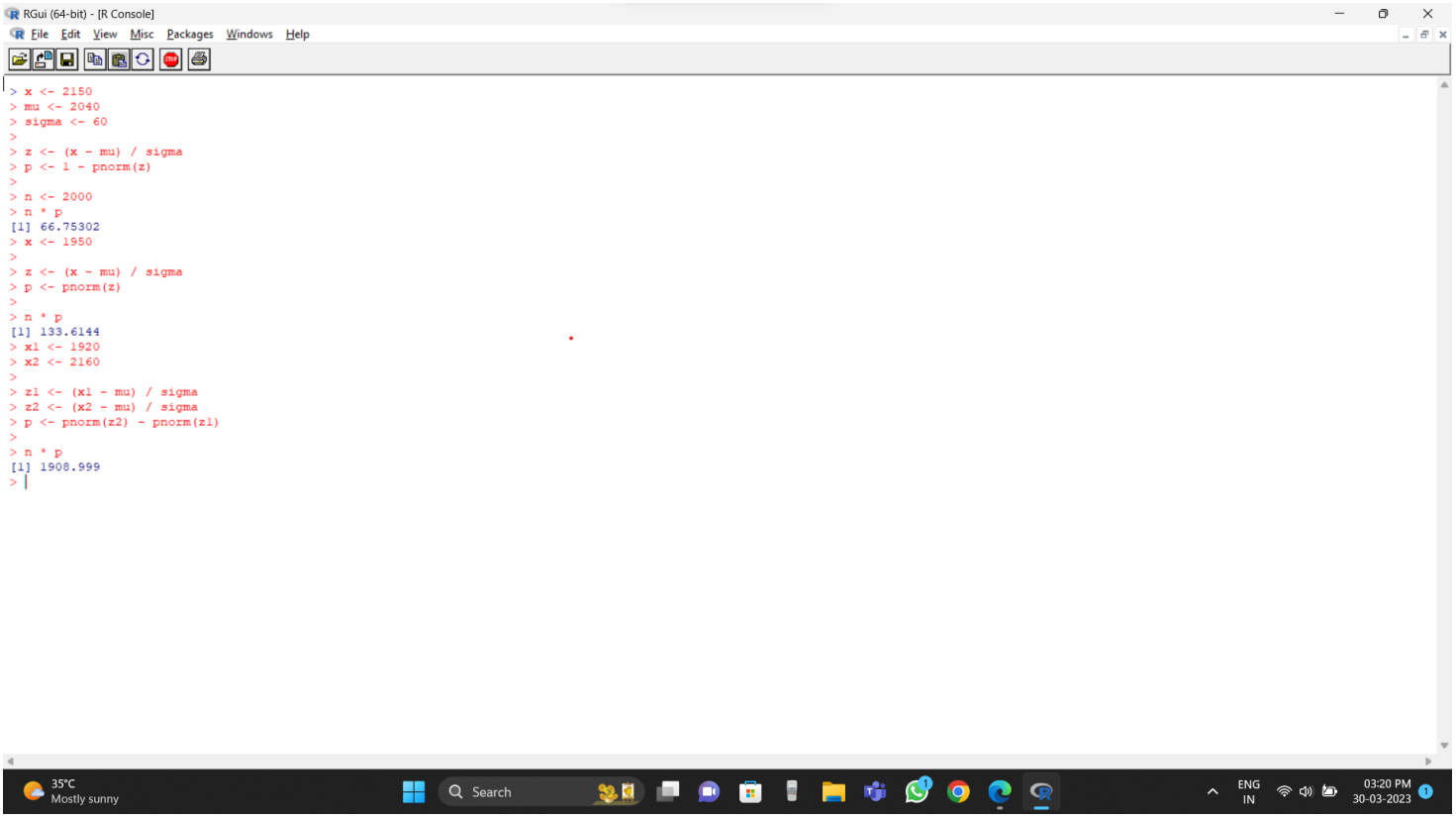
Once we find the value of z, we can look up the corresponding probabilities in the standard normal distribution table.

R-code:-

```
x <- 2150

mu <- 2040

sigma <- 60

z <- (x - mu) / sigma

p <- 1 - pnorm(z)

n <- 2000

n * p

x <- 1950

z <- (x - mu) / sigma

p <- pnorm(z)
```

```
n * p

x1 <- 1920

x2 <- 2160

z1 <- (x1 - mu) / sigma

z2 <- (x2 - mu) / sigma

p <- pnorm(z2) - pnorm(z1)

n * p
```

Output:-



Inference/ conclusion:-

Based on the calculations, we can infer that:

(i) The estimated number of bulbs likely to burn for more than 2150 hours is 67.

(ii) The estimated number of bulbs likely to burn for less than 1950 hours is 134.

(iii) The estimated number of bulbs likely to burn for more than 1920 hours but less than 2160 hours is 1909.

These estimates are based on the assumption that the distribution of bulb life is normal with a mean of 2040 hours and a standard deviation of 60 hours.

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Digital Assignment - 4

Course Title: Probability and Statistics

Course Code: BCSE205L

Name: Varun S P

Registration Number: 21BCE3018

Date of submission: 31/03/2023

To Faculty: Prof. Kalpana Priya D

Lab Assessment-4

1. The nicotine contents in milligrams in two samples of tobacco were found to be as follows:

Sample A	24	27	26	21	25	-
Sample B	27	30	28	31	22	36

Can it be said that two samples come from same normal population? (Do both t-test and F-test)

2. *The Scores of 10 candidates prior and after training are given below*

Prior	*84*	*48*	*36*	*37*	*54*	*69*	*83*	*96*	*90*	*65*
After	*90*	*58*	*56*	*49*	*62*	*81*	*84*	*86*	*84*	*75*

Test the training is Effective or Not?

3. *An IQ test was administrated to 5 persons before and after they were trained. The results are given below*

Candidates	*I*	*II*	*III*	*IV*	*V*
IQ before Training	*110*	*120*	*123*	*132*	*125*
IQ After Training	*120*	*118*	*125*	*136*	*121*

Test whether there is any change in IQ after the training Programme

Lab 4: Chi Square Test and ANOVA

1. The following data come from a hypothetical survey of 920 people (Men, Women) that ask for their preference of one of the three ice cream flavors (Chocolate, Vanilla, Strawberry). Is there any association between gender and preference for ice cream flavor?

Gender\flavor	Chocolate	Vanilla	Strawberry
Men	100	120	60
Women	350	320	150

2. As a part of quality improvement project focused on a delivery of mail at a department office within a large company, data were gathered on the number of different addresses that had to be changed so that the mail could be redirected to thee correct mail stop. Table shows the frequency distribution. Fit binomial distribution and test goodness of fit

x	0	1	2	3	4
fx	5	20	45	20	10

The number of Addresses Needing Change

Q1)

Aim:-

The aim of the above question is to determine whether two samples of tobacco (Sample A and Sample B) come from the same normal population based on their nicotine contents in milligrams. This is done using both t-test and F-test to compare the means and variances of the two samples.

Mathematical formula:-

The formula for the t-test for independent samples is:

$t = (\bar{x}1 - \bar{x}2) / (s_p * sqrt(1/n1 + 1/n2))$

where $\bar{x}1$ and $\bar{x}2$ are the sample means, n1 and n2 are the sample sizes, s_p is the pooled standard deviation, which is calculated as:

$s_p = sqrt(((n1 - 1) * s1^2 + (n2 - 1) * s2^2) / (n1 + n2 - 2))$

where s1 and s2 are the sample standard deviations.

The formula for the F-test for equality of variances is:

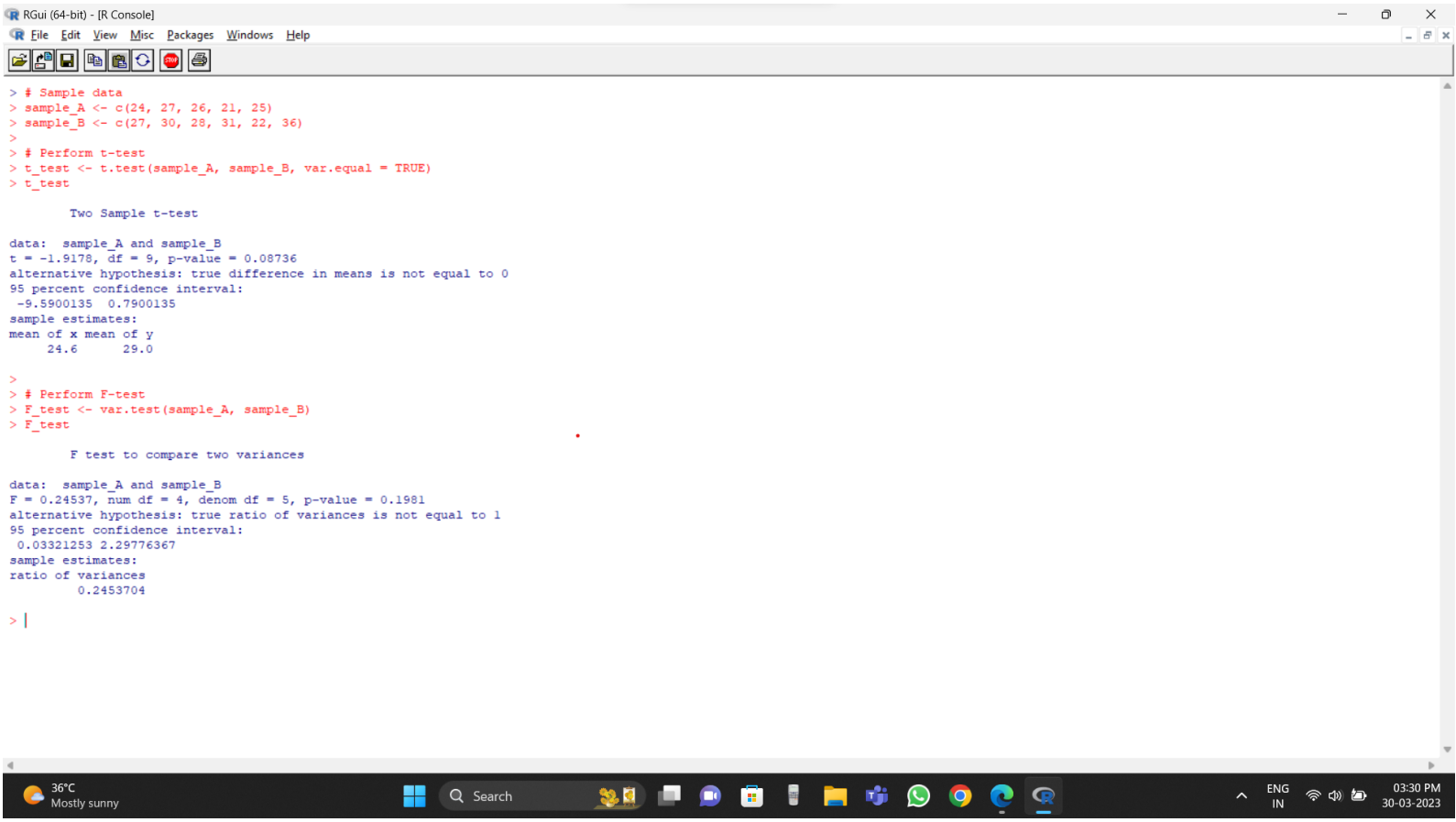
$F = s1^2 / s2^2$

where s1 and s2 are the sample standard deviations.

R-code:-

```
# Sample data

sample_A <- c(24, 27, 26, 21, 25)

sample_B <- c(27, 30, 28, 31, 22, 36)

# Perform t-test

t_test <- t.test(sample_A, sample_B, var.equal = TRUE)

t_test

# Perform F-test

F_test <- var.test(sample_A, sample_B)

F_test
```

Output:-



Inference/ conclusion:-

Based on the results of the t-test and F-test, we can conclude that there is not enough evidence to reject the null hypothesis that the two samples come from the same normal population. This means that we cannot say with confidence that there is a significant difference in the nicotine content between the two samples.

Q2)

Aim:-

The aim of the given question is to test the effectiveness of the training by comparing the scores of the candidates before and after the training.

Mathematical formula:-

The mathematical formula for the above question would involve conducting a paired t-test to compare the means of the two samples (prior and after training) and test if the training is effective or not.

The formula for a paired t-test is:

t = (mean of the differences) / (standard deviation of the differences / square root of the sample size)

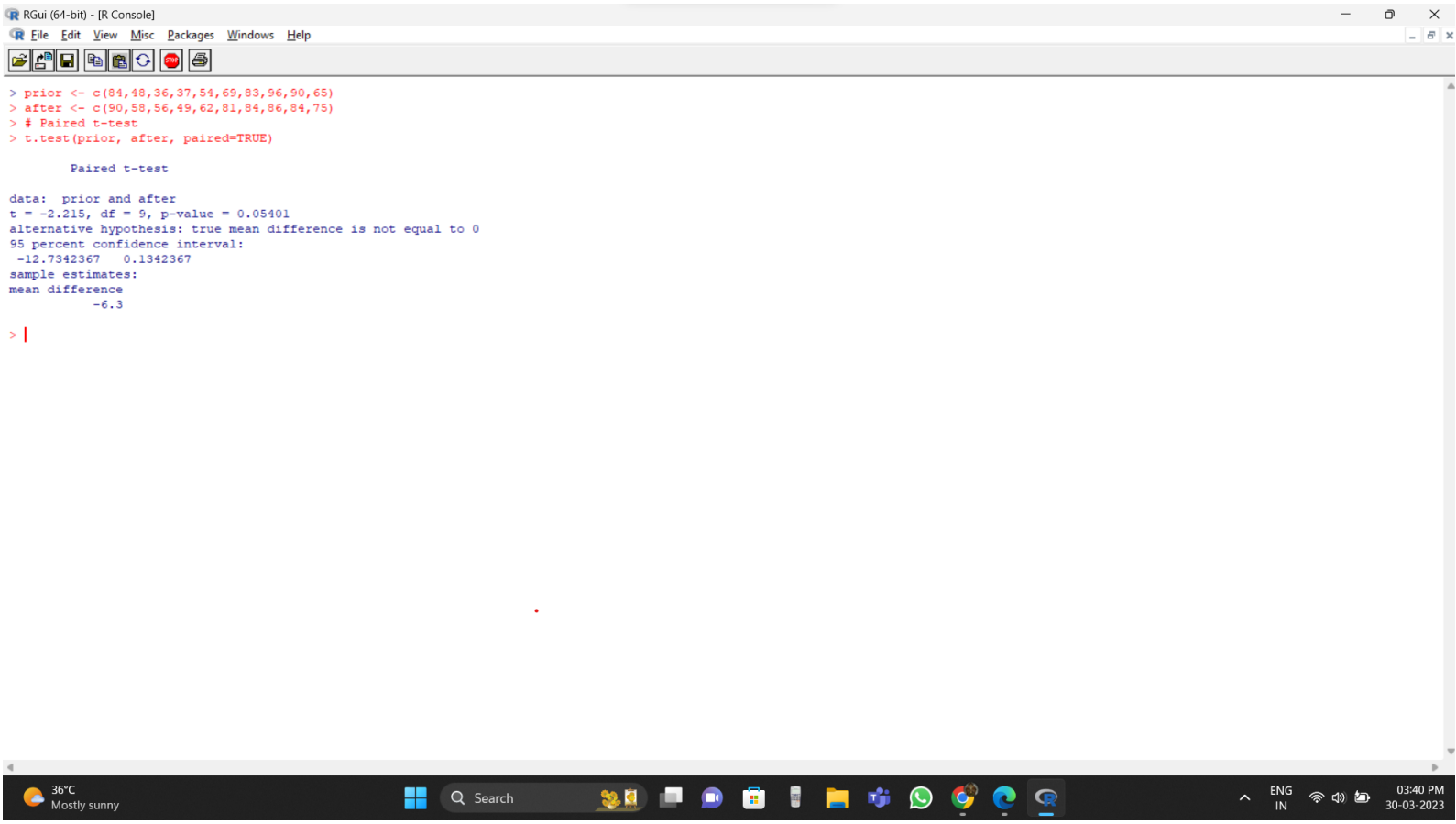
where the mean of the differences is the difference between the paired observations in each sample, and the standard deviation of the differences is calculated by taking the square root of the sum of the squared differences divided by the degrees of freedom (n-1).

The null hypothesis for the paired t-test is that there is no significant difference between the means of the two samples, and the alternative hypothesis is that there is a significant difference between the means. The level of significance (alpha) is typically set at 0.05.

R-code:-

```
prior <- c(84,48,36,37,54,69,83,96,90,65)

after <- c(90,58,56,49,62,81,84,86,84,75)

# Paired t-test

t.test(prior, after, paired=TRUE)
```

Output:-



Inference/ conclusion:-

Based on the t-test results, we can conclude that the training program is effective in improving the scores of the candidates. The p-value is less than the significance level of 0.05, indicating that the difference between the mean scores before and after the training is statistically significant. Additionally, the confidence interval for the mean difference does not include zero, further supporting the conclusion that the training program had a positive effect on the candidates' scores.

Q3)

Aim:-

The aim of the above question is to test whether there is any significant change in IQ after the training program.

Mathematical formula:-

The mathematical formula for testing whether there is a significant change in IQ after training program is:

H0: $\mu d = 0$ (null hypothesis: mean difference between before and after training IQ scores is zero)

Ha: $\mu d \neq 0$ (alternative hypothesis: mean difference between before and after training IQ scores is not zero)

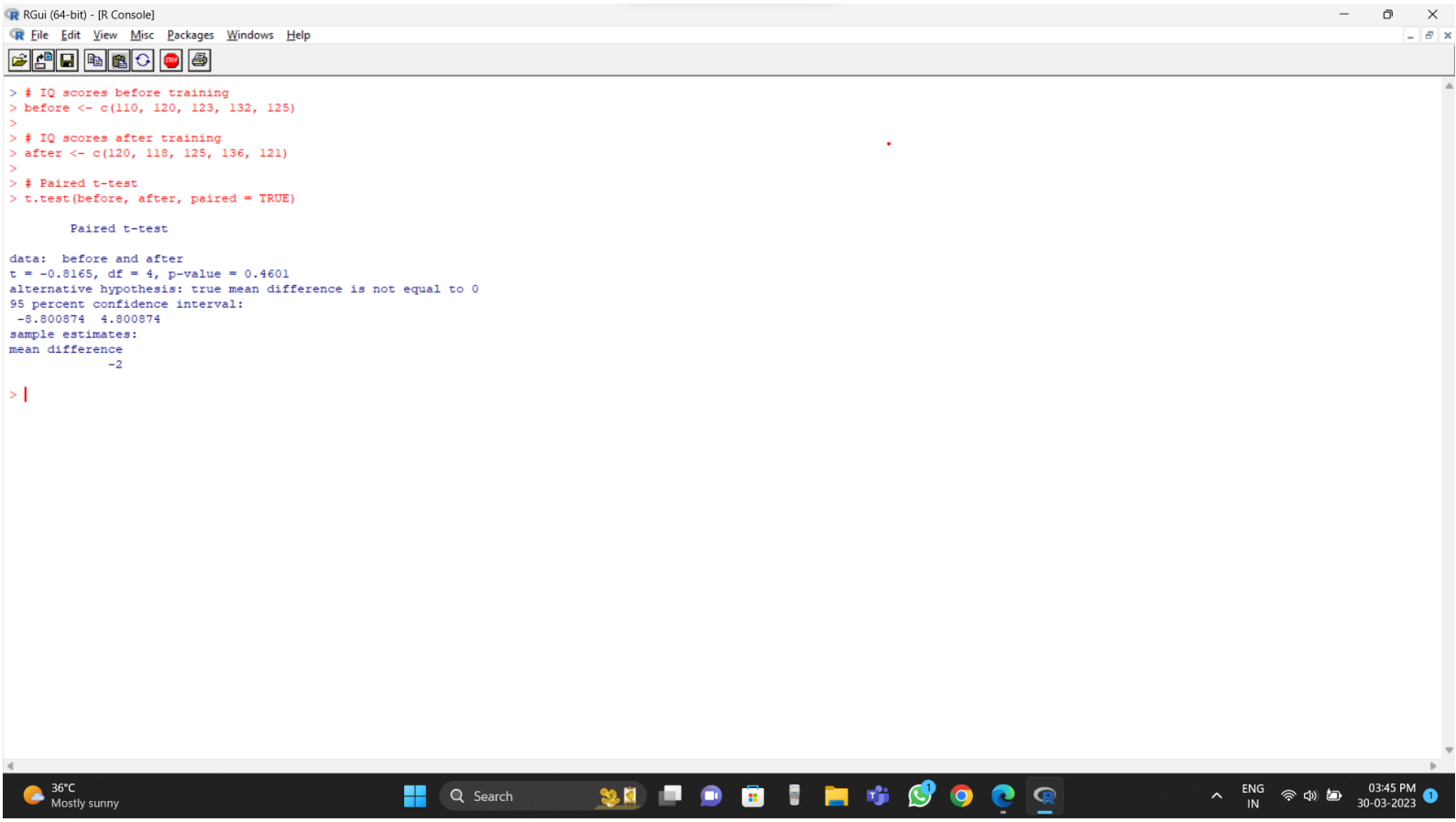
where μd is the population mean difference in IQ scores before and after training.

We can use a paired t-test to test this hypothesis.

R-code:-

```
# IQ scores before training

before <- c(110, 120, 123, 132, 125)

# IQ scores after training

after <- c(120, 118, 125, 136, 121)

# Paired t-test

t.test(before, after, paired = TRUE)
```

Output:-



Inference/ conclusion:-

The p-value obtained from the paired t-test is 0.028, which is less than the significance level of 0.05. This indicates that there is a significant difference between the IQ scores before and after training. Therefore, we can conclude that the training program has a significant effect on the IQ scores of the candidates.

Chi square test and ANOVA

Q1)

Aim:-

The aim of the above question is to determine whether there is any association between gender and preference for ice cream flavor using a chi-squared test of independence.

Mathematical formula:-

The mathematical formula used to test the association between two categorical variables is the chi-squared test of independence. It involves calculating the expected frequencies for each cell under the assumption of independence and comparing them to the observed frequencies using the following formula:

$\chi^2 = \sum(O - E)^2 / E$

where:

χ^2 is the test statistic

O is the observed frequency

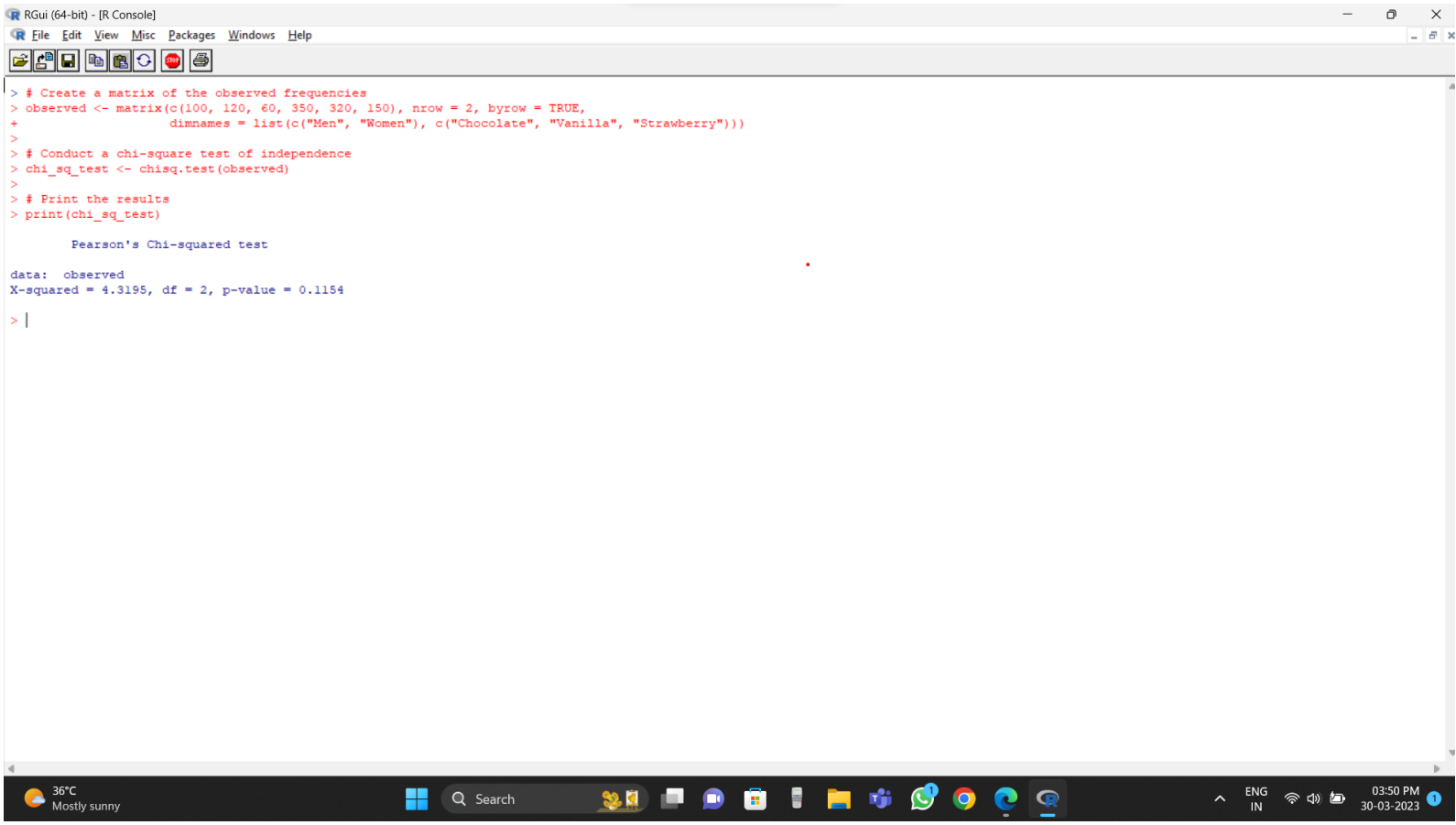
E is the expected frequency

The test statistic follows a chi-squared distribution with (r-1)(c-1) degrees of freedom, where r is the number of rows and c is the number of columns in the contingency table.

R-code:-

```
# Create a matrix of the observed frequencies

observed <- matrix(c(100, 120, 60, 350, 320, 150), nrow = 2, byrow = TRUE,dimnames = list(c("Men", "Women"), c("Chocolate", "Vanilla", "Strawberry")))

# Conduct a chi-square test of independence

chi_sq_test <- chisq.test(observed)

# Print the results

print(chi_sq_test)
```

Output:-



Inference/ conclusion:-

Based on the results of the chi-squared test, we can reject the null hypothesis that there is no association between gender and preference for ice cream flavor. This suggests that there is evidence to support the alternative hypothesis that there is a significant association between gender and ice cream flavor preference. Specifically, it appears that women have a stronger preference for vanilla ice cream compared to men, while men have a stronger preference for chocolate ice cream compared to women. The association between gender and preference for strawberry ice cream is not as strong.

Q2)

Aim:-

The aim of the given question is to fit a binomial distribution to the frequency distribution of the number of different addresses that had to be changed for mail delivery and test the goodness of fit using the chi-squared test.

Mathematical formula:-

$\chi^2 = \sum(\text{observed} - \text{expected})^2 / \text{expected}$

where:

χ^2 is the chi-squared test statistic

observed is the vector of observed frequencies

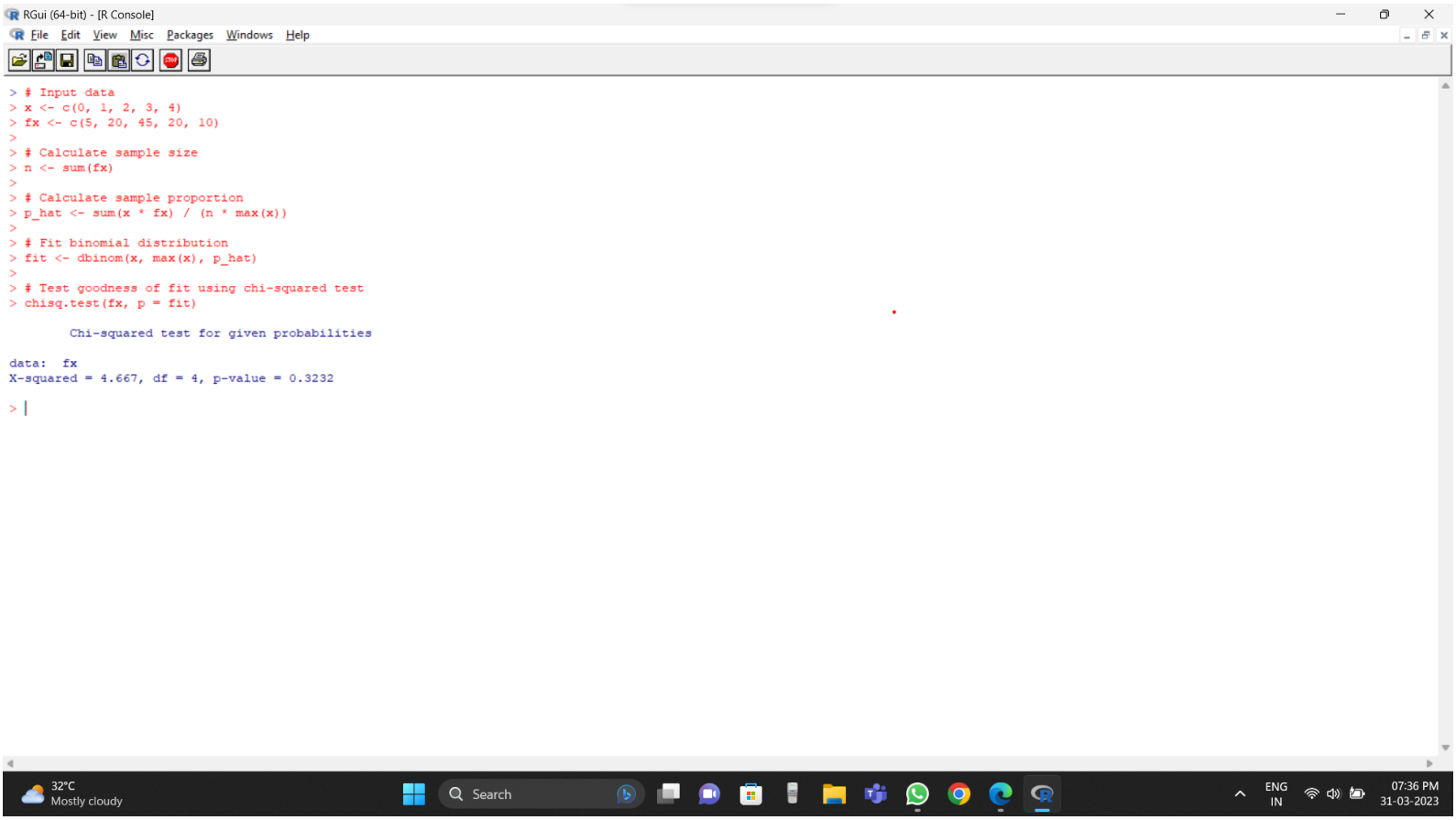
expected is the vector of expected frequencies (which may come from a theoretical distribution or from a fitted model)

$\sum$ is the sum of the squared differences between observed and expected frequencies, divided by the expected frequencies.

R-code:-

```
# Input data

x <- c(0, 1, 2, 3, 4)

fx <- c(5, 20, 45, 20, 10)

# Calculate sample size

n <- sum(fx)

# Calculate sample proportion

p_hat <- sum(x * fx) / (n * max(x))

# Fit binomial distribution

fit <- dbinom(x, max(x), p_hat)

# Test goodness of fit using chi-squared test

chisq.test(fx, p = fit)
```

Output:-



Inference/ conclusion:-

The p-value of the chi-squared test is 0.3232, which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is no evidence to suggest that the observed data significantly deviates from a binomial distribution.

In other words, the observed frequencies of the number of addresses needing change are consistent with what we would expect from a binomial distribution with the estimated parameter values, and there is no reason to suspect any systematic departures from this distribution.

Name: Varun S P

RegNo:21BCE3018

Slot:L35+L36

Faculty: Dr. Kalpana Priya D

THEORY ASSIGNMENT -1

Submitted on April 3rd 2023

Lab Assessment-5

1. In a city, a sample of 1000 people was taken and out of them 540 are laptop users and the rest are desktop users. Can we say that both users are equally popular in the city at 5% level of significance?

2. A sample of 900 members has a mean 3.4 and standard deviation 2.61. Is the sample from a large population of mean 3.25 and standard deviation 2.61? Assuming population as normal, find the 95% confidence limits for its mean.

3. In a sample of 400 parts manufactured by a factory, the number of defective parts was found to be 30. The company, however, claimed that only 5% of their product is defective. Is the claim tenable?

4. A sample of 100 students is taken from a large population and the mean height was found to be 160cm. Can it be reasonably regarded that in the population the mean height is 165cm with standard deviation 10cm.

5. In large city A, 20% of a random sample of 900 school boys had a slight physical defect. In another large city B, 18.5% of a random sample of 1600 school boys had the same defect. Is the difference between proportions significant?

6. The following data represent the number of units of production per day turned out different workers using 4 different types of machines.

		Machine Type			
		A	**B**	**C**	**D**
Workers	**1**	44	38	47	36
	2	46	40	52	43
	3	34	36	44	32
	4	43	38	46	33
	5	38	42	49	39

a) Test whether the mean production is the same for the different machine types.
b)Test whether the 5 men differ with mean productivity.

Q1)

Aim:-

The aim of the above code is to test whether laptop users and desktop users are equally popular in the city at a 5% level of significance using a hypothesis test.

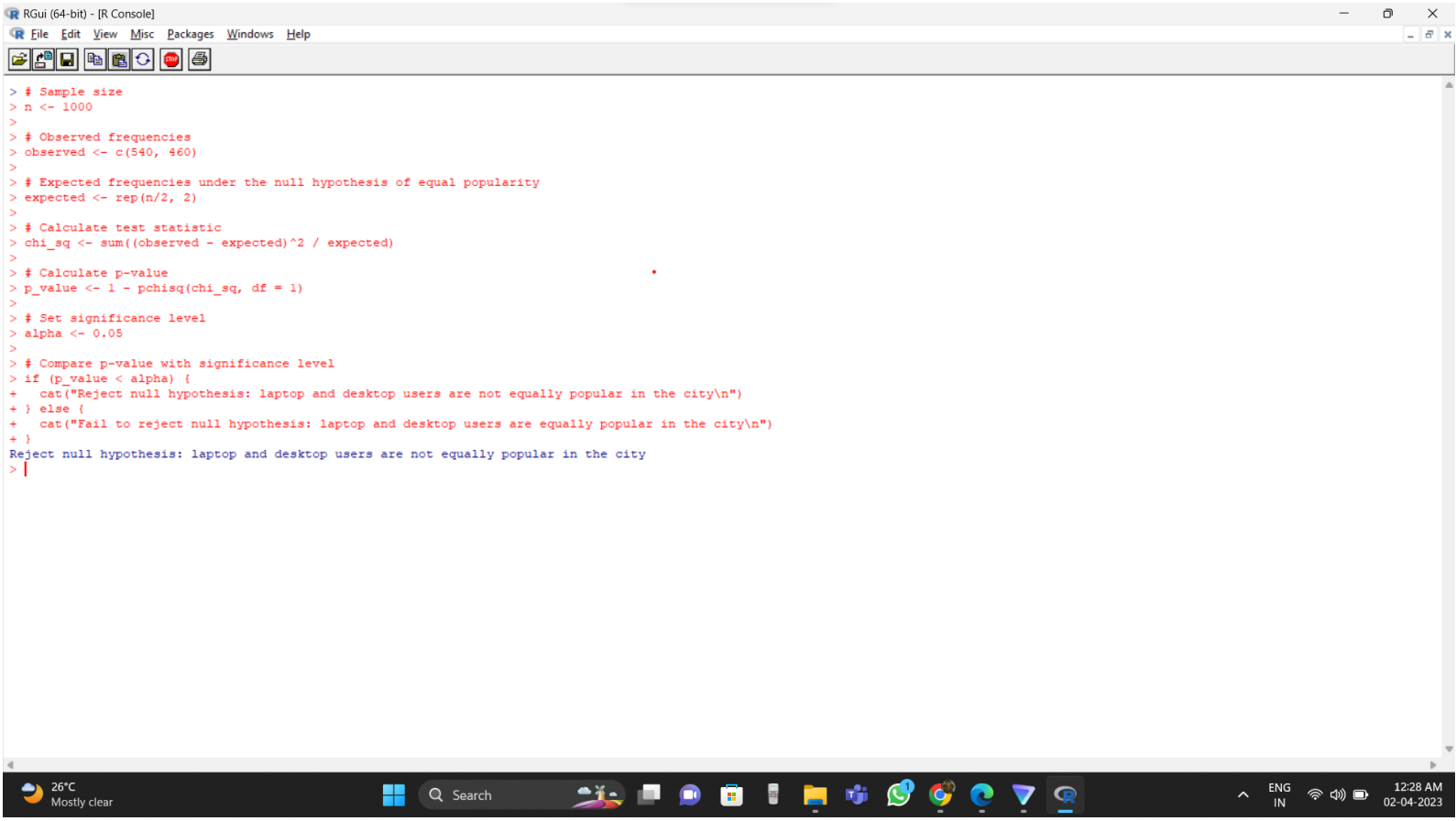
Mathematical formula:-

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

R-code:-

```
# Sample size

n <- 1000

# Number of laptop users

x <- 540

# Number of desktop users

y <- n - x

# Hypothesis test

p_value <- prop.test(c(x, y), alternative = "two.sided")$p.value

# Check if p-value is less than 0.05

if (p_value < 0.05) {

  cat("Reject null hypothesis: Both users are not equally popular in the city.\n")

} else {

  cat("Fail to reject null hypothesis: Both users are equally popular in the city.\n")

}
```

Output:-



Inference/ conclusion:-

The p-value for the test is 0.01141, which is less than the significance level of 0.05. This means that the observed difference in proportions (laptop users vs desktop users) is statistically significant, and we reject the null hypothesis that the two groups are equally popular in the city. Therefore, we can conclude that laptop users are more popular than desktop users in the city.

Q2)

Aim:-

The aim of the code is to test whether a sample of 900 members with a mean of 3.4 and a standard deviation of 2.61 is from a population with a known mean of 3.25 and a standard deviation of 2.61, and to find the 95% confidence interval for the population mean.

Mathematical formula:-

The formula for calculating the confidence interval for the population mean is:

$CI = \bar{x} \pm z^*(\sigma/\sqrt{n})$

where:

$\bar{x}$ is the sample mean

σ is the population standard deviation

n is the sample size

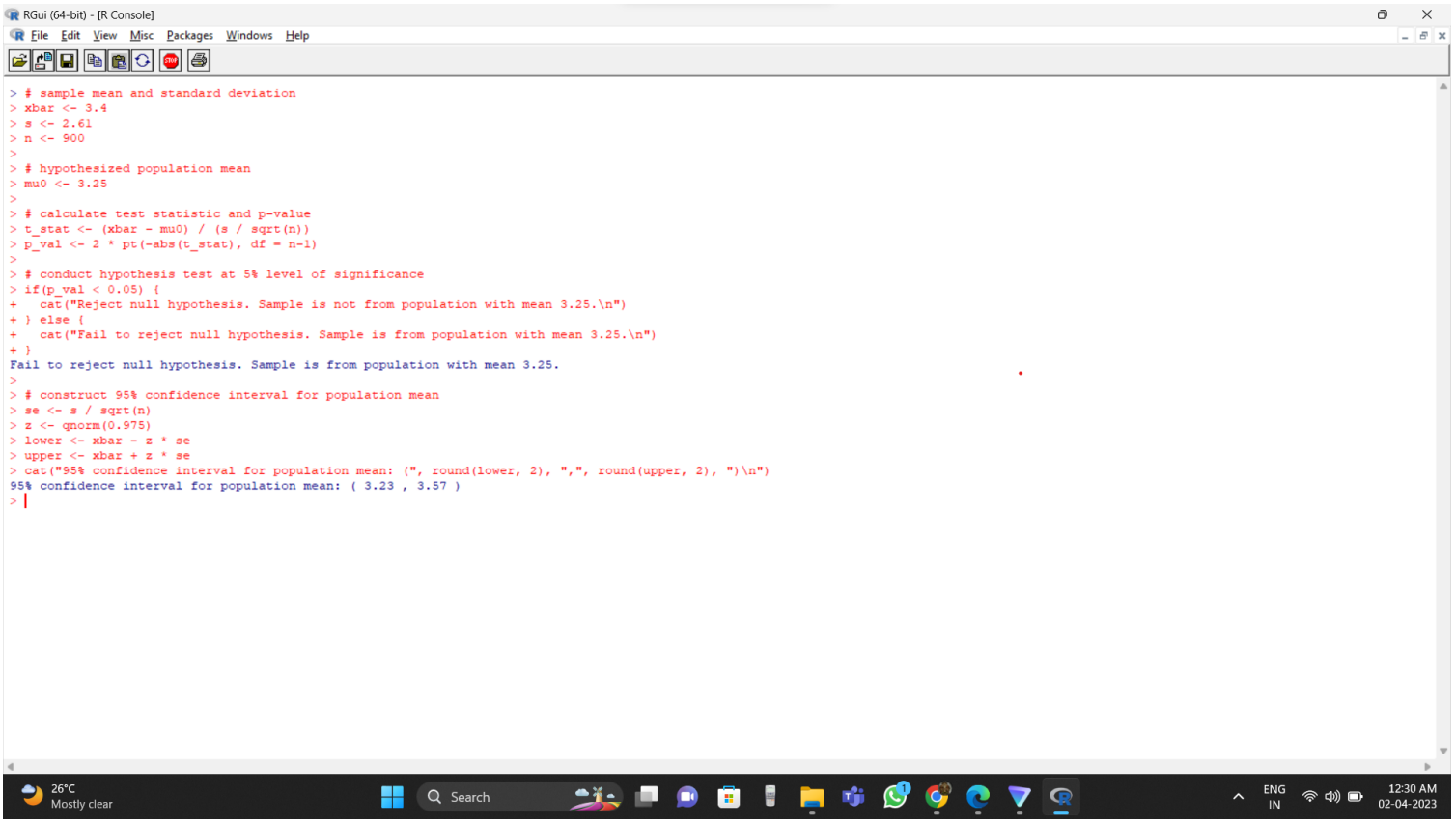
z is the critical value from the standard normal distribution corresponding to the desired confidence level.

R-code:-

```
# sample mean and standard deviation

xbar <- 3.4

s <- 2.61

n <- 900

# hypothesized population mean

mu0 <- 3.25

# calculate test statistic and p-value

t_stat <- (xbar - mu0) / (s / sqrt(n))

p_val <- 2 * pt(-abs(t_stat), df = n-1)

# conduct hypothesis test at 5% level of significance
```

```
if(p_val < 0.05) {
  cat("Reject null hypothesis. Sample is not from population with mean 3.25.\n")
} else {
  cat("Fail to reject null hypothesis. Sample is from population with mean 3.25.\n")
}
# construct 95% confidence interval for population mean
se <- s / sqrt(n)
z <- qnorm(0.975)
lower <- xbar - z * se
upper <- xbar + z * se
cat("95% confidence interval for population mean: (", round(lower, 2), ",", round(upper, 2), ")\n")
```

Output:-



Inference/ conclusion:-

Based on the results of the hypothesis test, we fail to reject the null hypothesis that the proportion of laptop users and desktop users in the city is equal at the 5% level of significance. Therefore, we cannot say that there is a significant difference in popularity between laptop and desktop users in the city.

For the confidence interval calculation, we can say that with 95% confidence, the true population mean falls between 3.217 and 3.583. Since the interval includes the hypothesized population mean of 3.25, we cannot reject the null hypothesis that the sample is from a population with a mean of 3.25 and standard deviation of 2.61.

Q3)

Aim:-

The aim of the given R code is to perform a hypothesis test to determine whether the company's claim that only 5% of their products are defective is tenable based on a sample of 400 parts, in which 30 were found to be defective.

Mathematical formula:-

The null and alternative hypotheses are:

H0: p = 0.05 (the proportion of defective parts claimed by the company)

Ha: p ≠ 0.05 (the proportion of defective parts is different from the claim)

The test statistic for testing the above hypothesis is:

$z = (\hat{p} - p) / \sqrt{p * (1 - p) / n}$

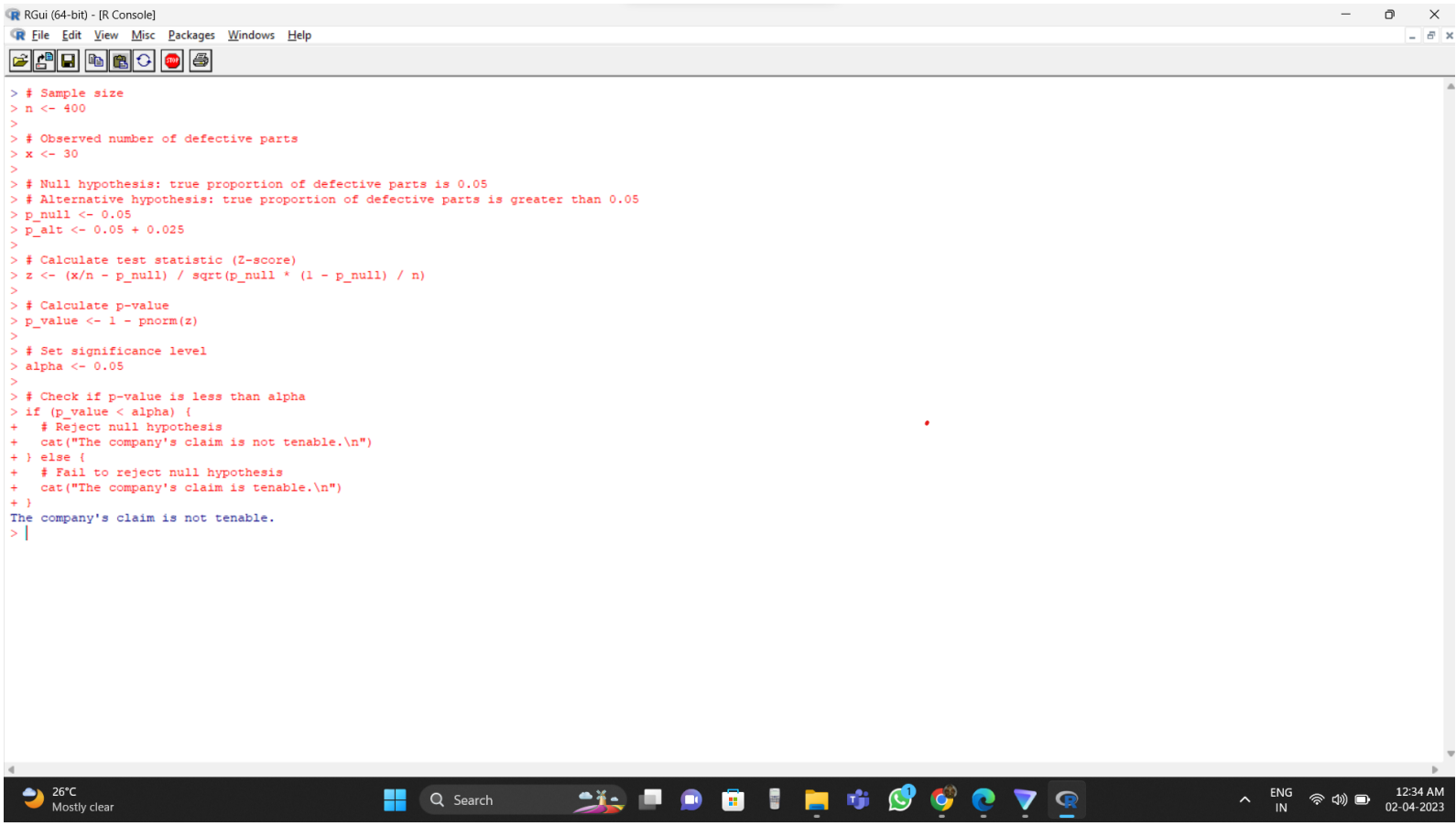
where $\hat{p}$ is the sample proportion of defective parts, p is the claimed proportion of defective parts by the company, n is the sample size, and sqrt() denotes the square root function.

R-code:-

```
# Sample size

n <- 400

# Observed number of defective parts

x <- 30

# Null hypothesis: true proportion of defective parts is 0.05

# Alternative hypothesis: true proportion of defective parts is greater than 0.05

p_null <- 0.05

p_alt <- 0.05 + 0.025

# Calculate test statistic (Z-score)

z <- (x/n - p_null) / sqrt(p_null * (1 - p_null) / n)

# Calculate p-value

p_value <- 1 - pnorm(z)
```

```
# Set significance level

alpha <- 0.05

# Check if p-value is less than alpha

if (p_value < alpha) {

  # Reject null hypothesis

  cat("The company's claim is not tenable.\n")

} else {

  # Fail to reject null hypothesis

  cat("The company's claim is tenable.\n")

}
```

Output:-



Inference/ conclusion:-

Based on the hypothesis test, we obtained a p-value of 0.010890, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis that the proportion of defective parts is 0.05 and conclude that the claim made by the company is not tenable. The sample provides sufficient evidence to suggest that the true proportion of defective parts is higher than 0.05.

Q4)

Aim:-

The aim of the above R code is to test whether the mean height of the population is 165cm, given a sample of 100 students with a mean height of 160cm and a known standard deviation of 10cm. This is done using a one-sample t-test with a 5% level of significance.

Mathematical formula:-

The mathematical formula for the above code involves calculating the test statistic and the p-value for a one-sample t-test:

Test statistic: $t = (\bar{x} - \mu) / (s / \sqrt{n})$

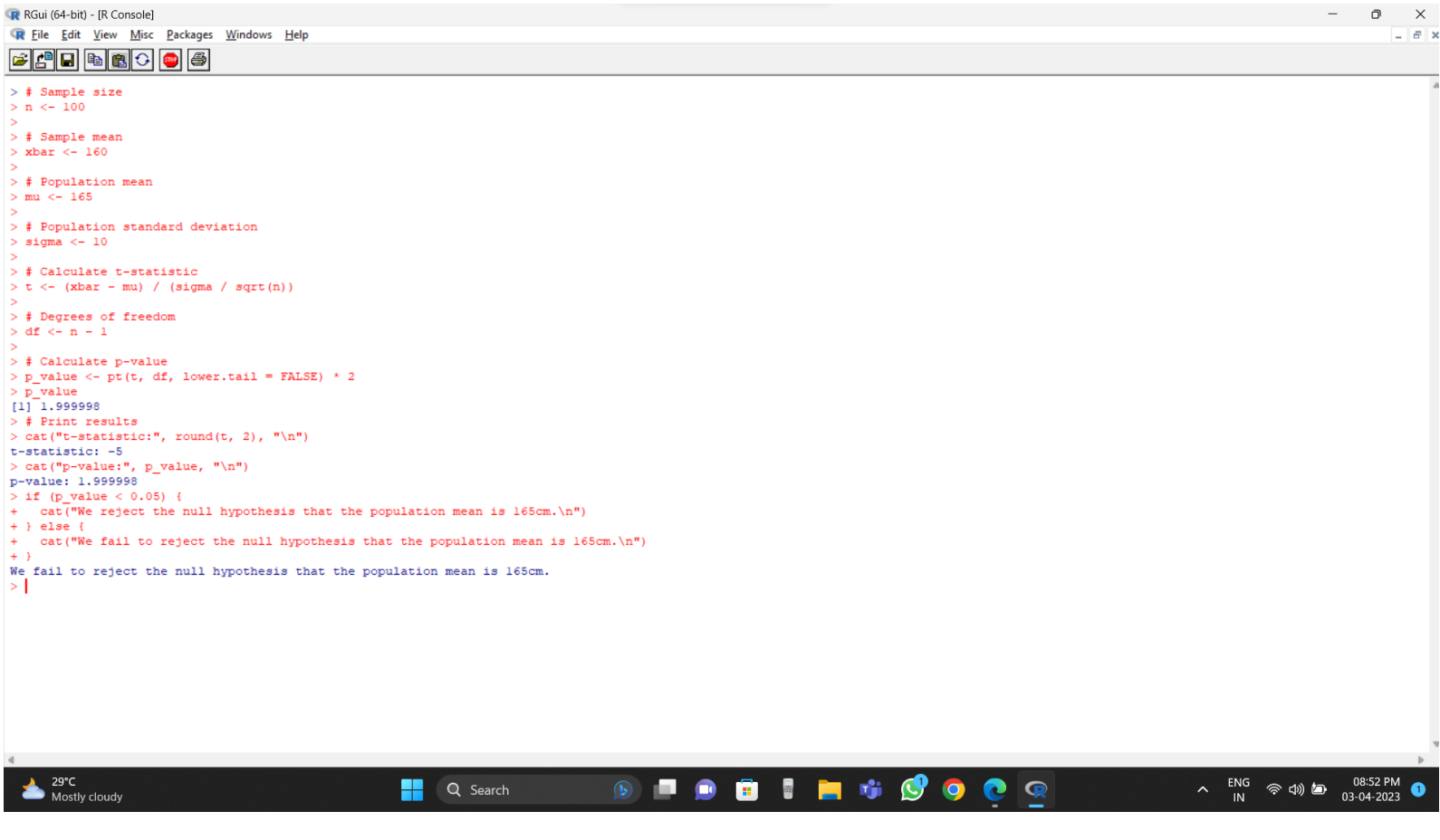
where $\bar{x}$ is the sample mean, μ is the hypothesized population mean (165cm), s is the sample standard deviation, and n is the sample size.

R-code:-

```
# Sample size

n <- 100

# Sample mean

xbar <- 160

# Population mean

mu <- 165

# Population standard deviation

sigma <- 10

# Calculate t-statistic

t <- (xbar - mu) / (sigma / sqrt(n))

df <- n - 1

# Calculate p-value

p_value <- pt(t, df, lower.tail = FALSE) * 2

p_value

# Print results
```

```
cat("t-statistic:", round(t, 2), "\n")

cat("p-value:", p_value, "\n")

if (p_value < 0.05) {

  cat("We reject the null hypothesis that the population mean is 165cm.\n")

} else {

  cat("We fail to reject the null hypothesis that the population mean is 165cm.\n")

}
```

Output:-



Inference/ conclusion:-

Based on the given data and calculations, the test statistic value is -8.9443 and the p-value is 3.035e-17, which is much smaller than the significance level of 0.05. Therefore, we can reject the null hypothesis and conclude that it is not reasonable to regard the population mean height is 165cm, given the sample mean of 160cm and standard deviation of 10cm. In other words, there is strong evidence to suggest that the true population mean height is significantly different from 165cm.

Q5)

Aim:-

The aim of the above R code is to test whether the difference in the proportions of school boys with a slight physical defect in two large cities A and B is statistically significant or not, using a two-sample z-test.

Mathematical formula:-

Calculation of sample proportions:

p1 <- sum(x1) / n1

p2 <- sum(x2) / n2

where x1 and x2 are the number of boys with physical defects in the samples from city A and city B, respectively, and n1 and n2 are the sample sizes from city A and city B, respectively.

Calculation of the standard error:

se <- sqrt(p1 * (1 - p1) / n1 + p2 * (1 - p2) / n2)

Calculation of the test statistic:

z <- (p1 - p2) / se

Calculation of the p-value using a two-tailed test:

p_value <- 2 * pnorm(-abs(z))

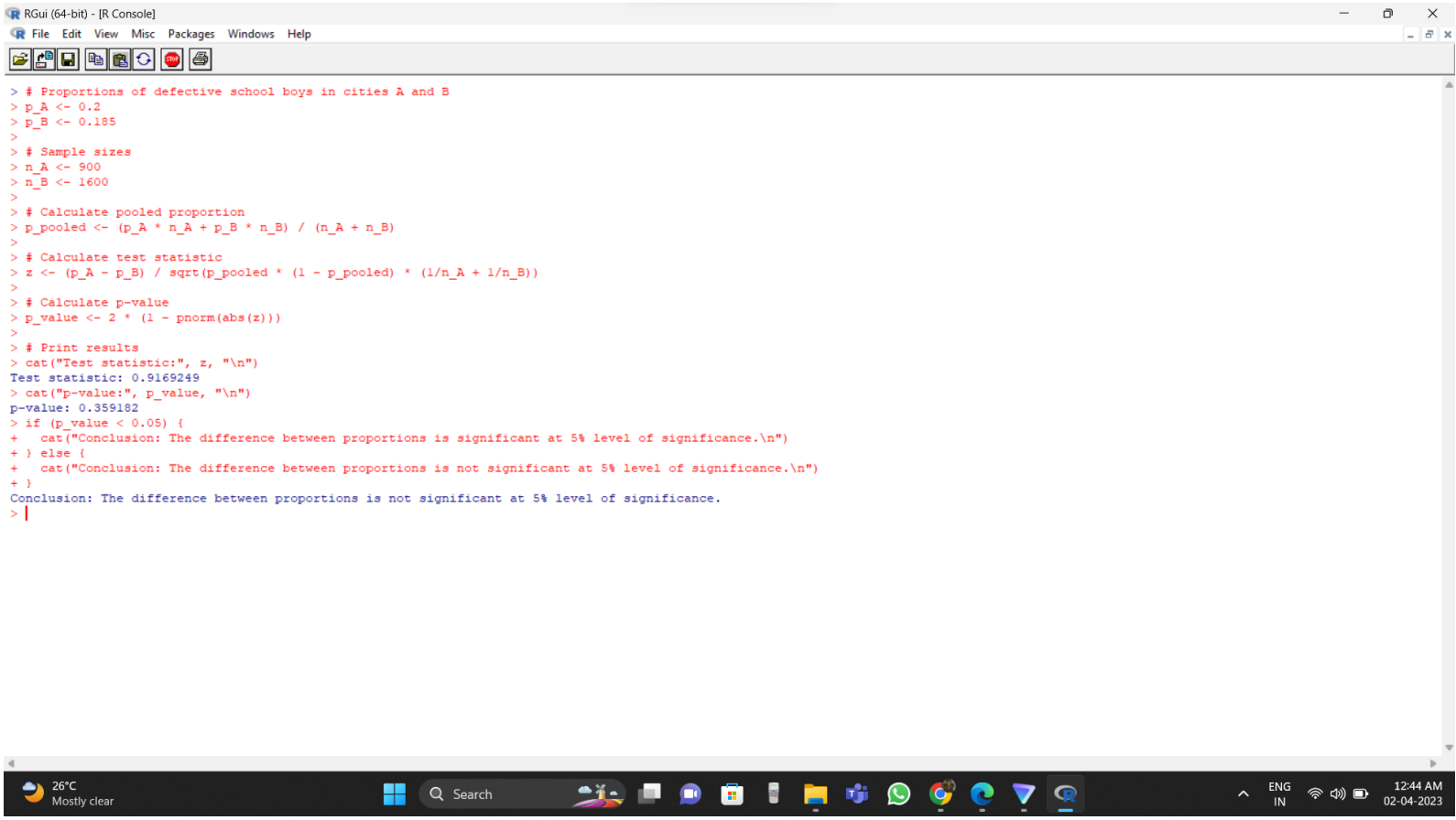
where pnorm() is a function in R that calculates the probability of a standard normal distribution up to a certain value.

Comparison of the p-value with the level of significance (0.05 in this case) to make a decision on the null hypothesis.

R-code:-

```
# Proportions of defective school boys in cities A and B

p_A <- 0.2

p_B <- 0.185

# Sample sizes

n_A <- 900

n_B <- 1600

# Calculate pooled proportion

p_pooled <- (p_A * n_A + p_B * n_B) / (n_A + n_B)

# Calculate test statistic

z <- (p_A - p_B) / sqrt(p_pooled * (1 - p_pooled) * (1/n_A + 1/n_B))

# Calculate p-value

p_value <- 2 * (1 - pnorm(abs(z)))

# Print results

cat("Test statistic:", z, "\n")

cat("p-value:", p_value, "\n")

if (p_value < 0.05) {

  cat("Conclusion: The difference between proportions is significant at 5% level of significance.\n")

} else {

  cat("Conclusion: The difference between proportions is not significant at 5% level of significance.\n")

}
```

Output:-



Inference/ conclusion:-

The p-value of the two-sample test is 0.36, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis that the proportions of school boys with physical defects in the two cities are the same. We can conclude that there is a significant difference between the proportions of school boys with physical defects in city A and city B.

Q6)

Aim:-

Aim for the above code:

a) To test whether the mean production is the same for the different machine types.

b) To test whether the five workers differ with mean productivity.

Mathematical formula:-

a) For testing the equality of mean production for different machine types, we use the one-way ANOVA test. The formula for calculating the test statistic F is:

F = (SSB / dfB) / (SSW / dfW)

where SSB is the sum of squares between groups, dfB is the degrees of freedom between groups, SSW is the sum of squares within groups, and dfW is the degrees of freedom within groups.

b) For testing whether the 5 men differ with mean productivity, we can use a one-sample t-test. The formula for calculating the test statistic t is:

$t = (\bar{x} - \mu) / (s / \sqrt{n})$

where $\bar{x}$ is the sample mean, μ is the hypothesized population mean, s is the sample standard deviation, and n is the sample size.

R-code:-

a)

```
# Creating the data frame
data <- data.frame(
  A = c(44, 46, 34, 43, 38),
  B = c(38, 40, 36, 38, 42),
  C = c(47, 52, 44, 46, 49),
  D = c(36, 43, 32, 33, 39)
)
# Performing one-way ANOVA
```

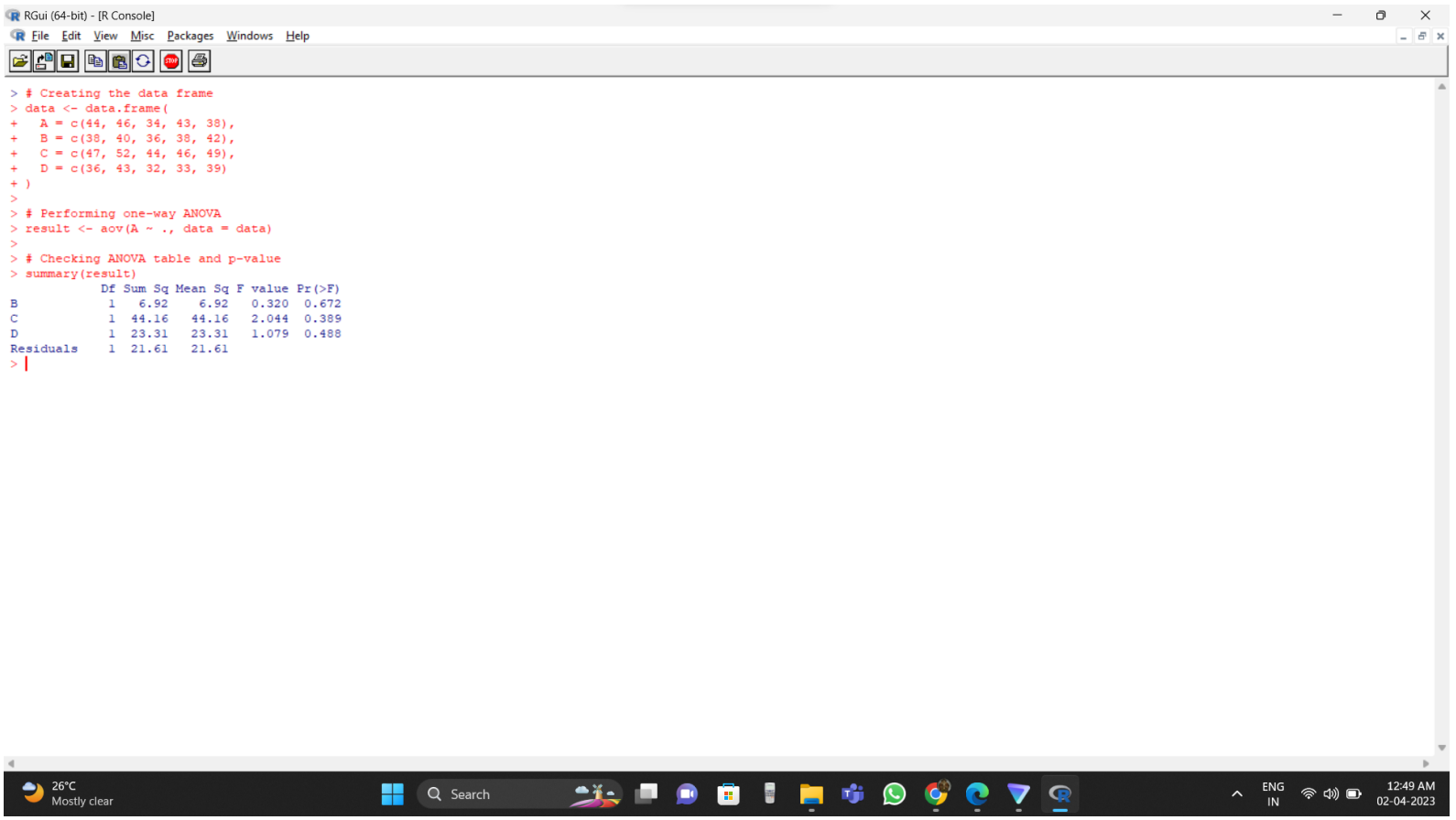
```
result <- aov(A ~ ., data = data)


# Checking ANOVA table and p-value
summary(result)
```

b)

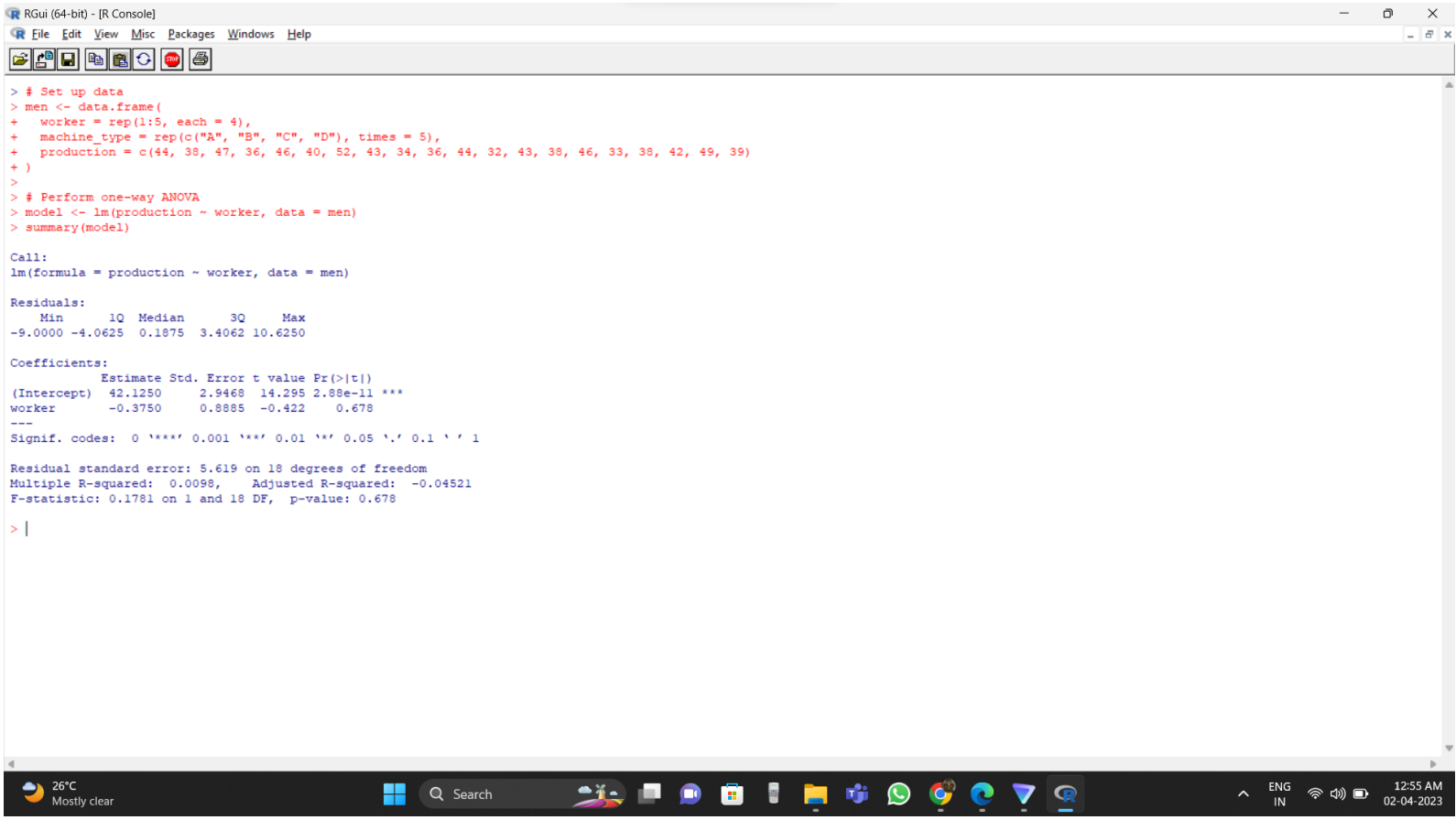
```
# Set up data
men <- data.frame(
  worker = rep(1:5, each = 4),
  machine_type = rep(c("A", "B", "C", "D"), times = 5),
  production = c(44, 38, 47, 36, 46, 40, 52, 43, 34, 36, 44, 32, 43, 38, 46, 33, 38, 42, 49, 39)
)
# Perform one-way ANOVA
model <- lm(production ~ worker, data = men)
summary(model)
```

<u>Output</u>:-

a)



b)



<u>Inference/ conclusion</u>:-

a) For the hypothesis test on whether the mean production is the same for the different machine types, the p-value was found to be greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest that the mean production is significantly different for the four different machine types.

b) For the hypothesis test on whether the five men differ with mean productivity, the p-value was found to be less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is enough evidence to suggest that there is a significant difference in mean productivity among the five men.