PES University, Bangalore Established under the Karnataka Act No. 16 of 2013

UE21CS342AA2 - Data Analytics , Worksheet 1 - ANOVA

Richa Shahi - shahiricha2412@gmail.com Abhay K Iyengar – abzee2002@gmail.com

Part 3: Analysis of Variance (ANOVA)

- Analysis of Variance (ANOVA) is a hypothesis testing procedure used for comparing means from several groups simultaneously.
- The objective of ANOVA is to check simultaneously whether population mean from more than two populations are different. ANOVA determines whether three or more populations are statistically different from each other.
- In a one-way ANOVA, we test whether the mean values of an outcome variable for different levels of a factor are different. Using multiple two sample t-tests to simultaneously test group means will result in incorrect estimation of Type-I error; ANOVA overcomes this problem.
- In two-way ANOVA, we check the impact of more than one factor simultaneously on several groups.

About the Dataset

The management at St. Clare's Primary School are concerned about the health of the students in the post Covid world. So they planned to introduce 4 different fitness routines labelled as A, B, C and D. These fitness plans include changes in diet, exercises and sleep routines. Students were randomly allocated to one of the fitness plans.

A, B, C and D are four different fitness plans introduced in the school.

You can download the datasheet from here.

- The table has four columns A, B, C and D which corresponds to the 4 different fitness routines.
- Each observations is the score obtained by the student out of 100 in the final exams.

We are going to analyze the affect of different fitness routines on the scores obtained by the student.

```r
# Install and load necessary packages
if (!requireNamespace("tidyverse", quietly = TRUE)) {
  install.packages("tidyverse")
}
if (!requireNamespace("moments", quietly = TRUE)) {
  install.packages("moments")
}

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(moments)
library(ggplot2)
```

Problem 1

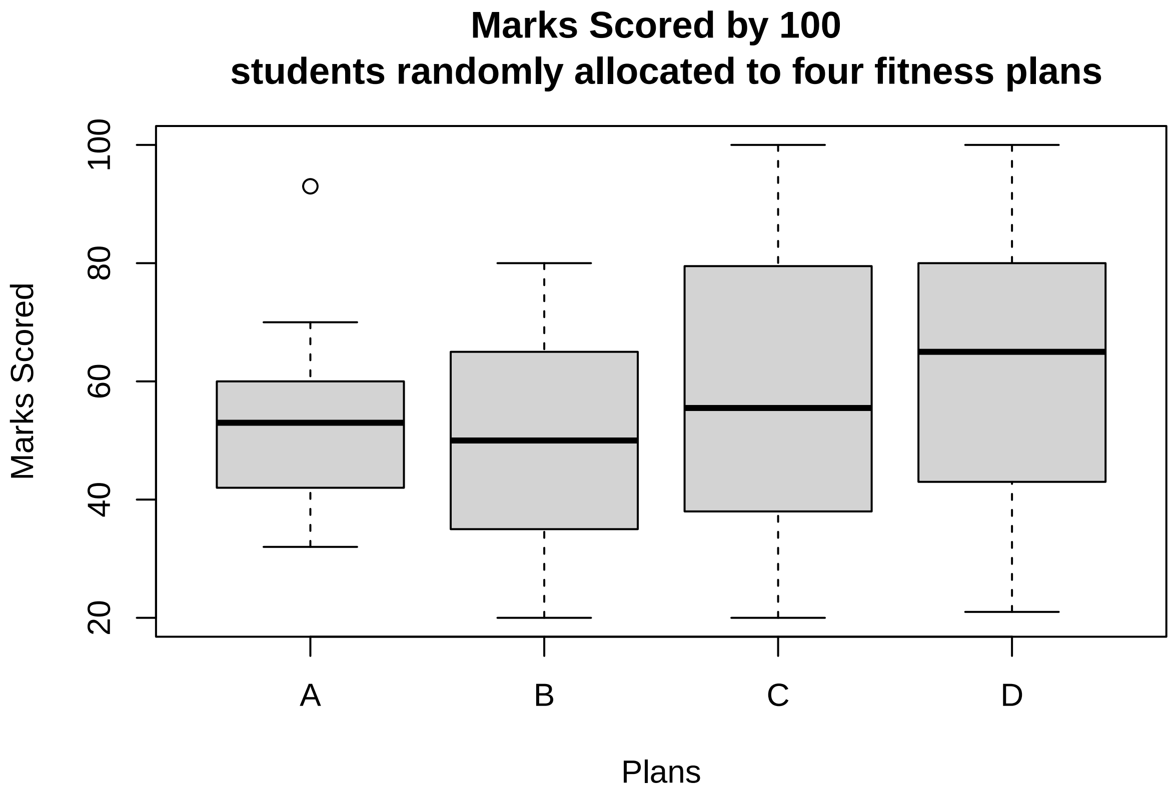
Read the data set and display the box plot for each of the fitness plans A, B, C, D. Analyze the box plot for outliers.

SOLUTION:

Read the data set and display the box plot for each of the fitness plans A, B, c, D

```r
# Read the CSV
data <- read.csv("students.csv", header=TRUE)
```

```r
# Create a box plot for all the fitness plans
boxplot(data$A, data$B, data$C, data$D, names = c("A", "B", "C", "D"),
        main = "Marks Scored by 100 \n students randomly allocated to four fitness plans",
        xlab = "Plans", ylab = "Marks Scored")
```



Only plan A has an outlier.

Problem 2

Is the data symmetrical or skewed for each group? Verify the normality assumption for ANOVA. (*Hint: Find the Pearson's moment coefficient of skewness and justify it with probability distribution function plot or you can also plot the Q-Q plot*)

SOLUTION: Finding the Pearson's moment coefficient of skewness to measure skew in each group.

```r
# Calculate skewness for each column
skewness_values <- sapply(data, skewness)
print(skewness_values)
```

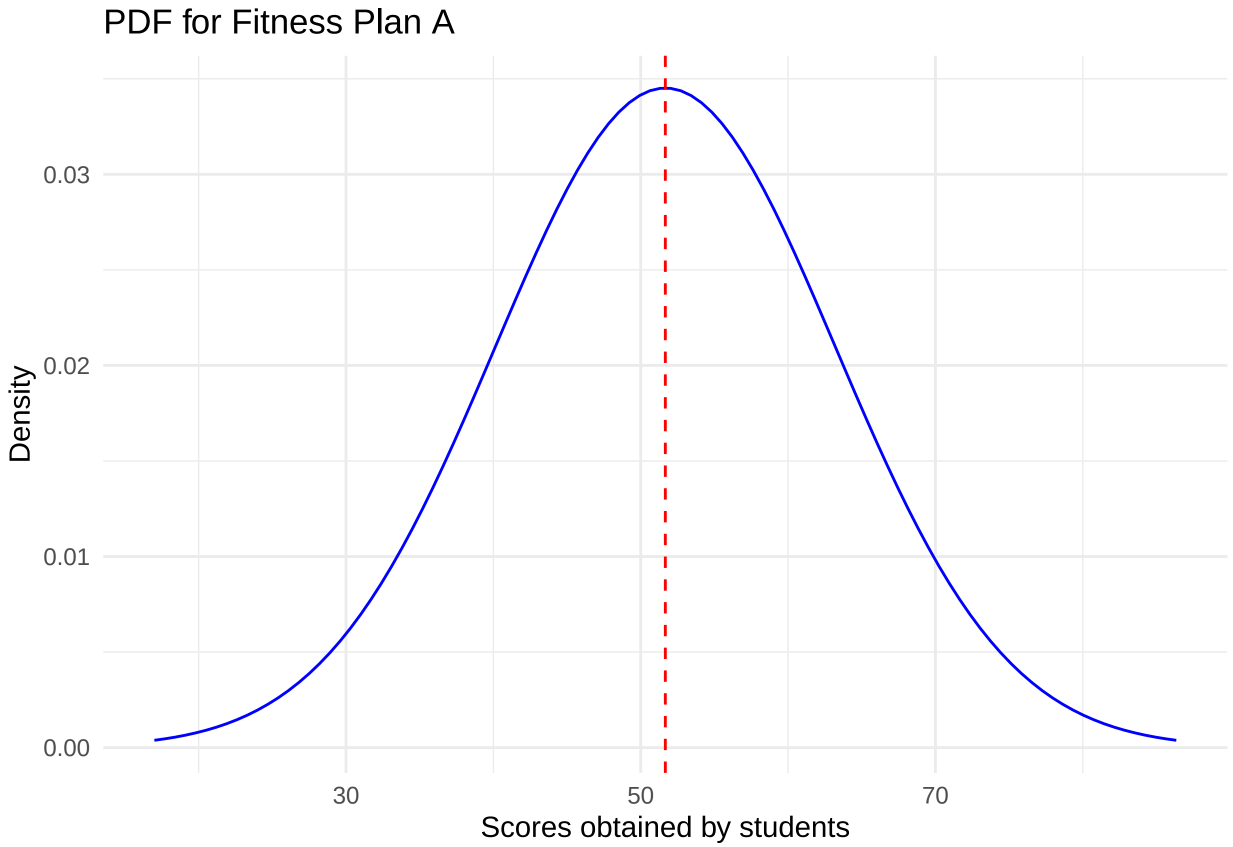
```
##           A           B           C           D
##  0.29487497  0.00960691  0.03168209 -0.06082271
```

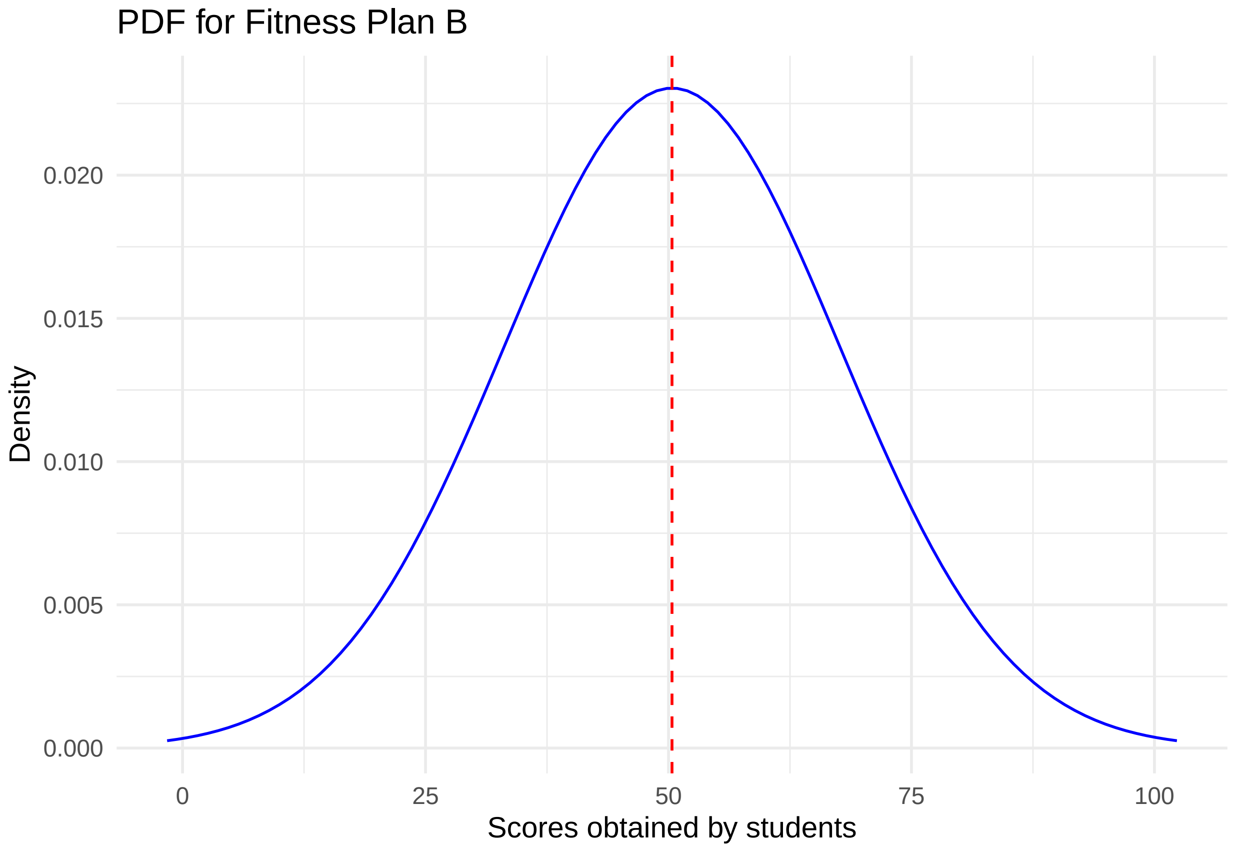
Columns A, B, C and D have the Pearson's moment coefficient of skewness in the range (-1,1) hence they can be accepted as symmetrical.

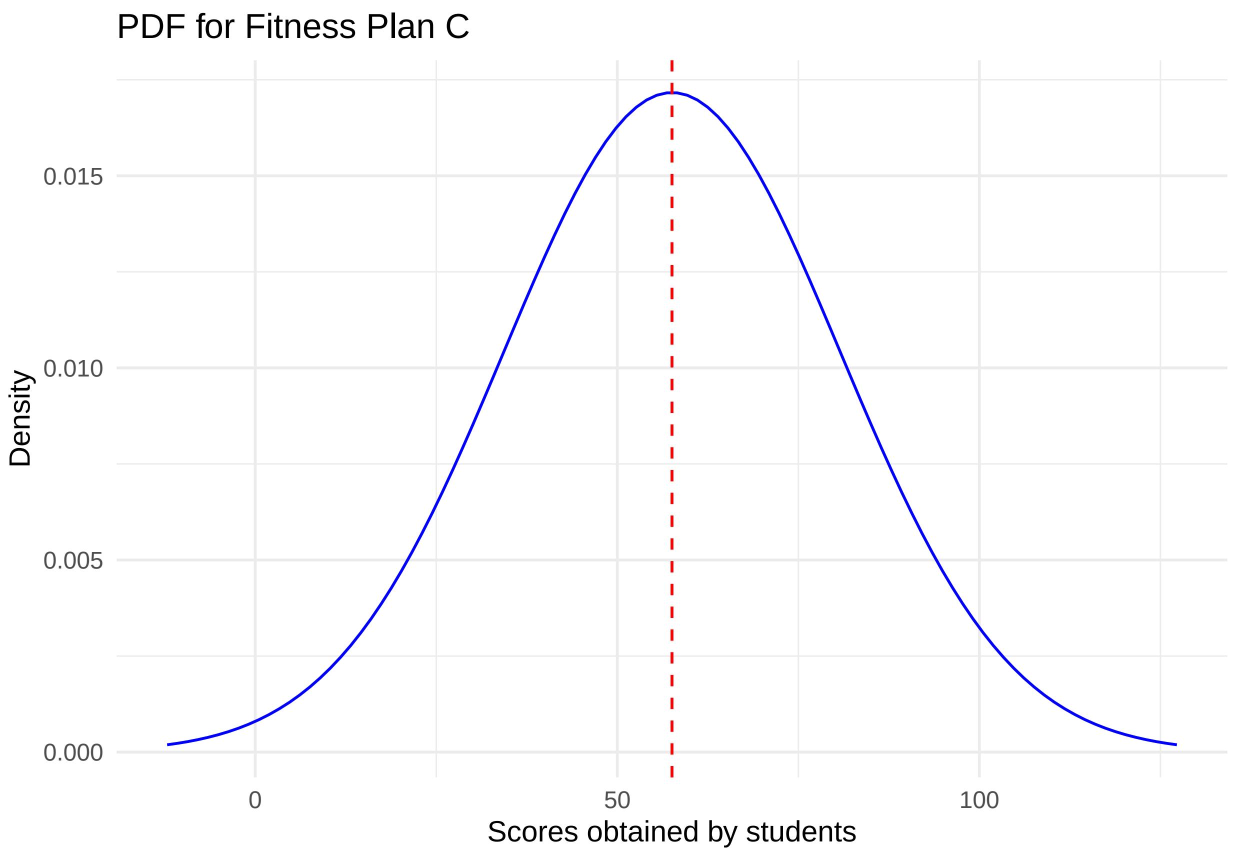
Because scores obtained is a continuous random variable let us also plot the probability density function to visualize the skew.

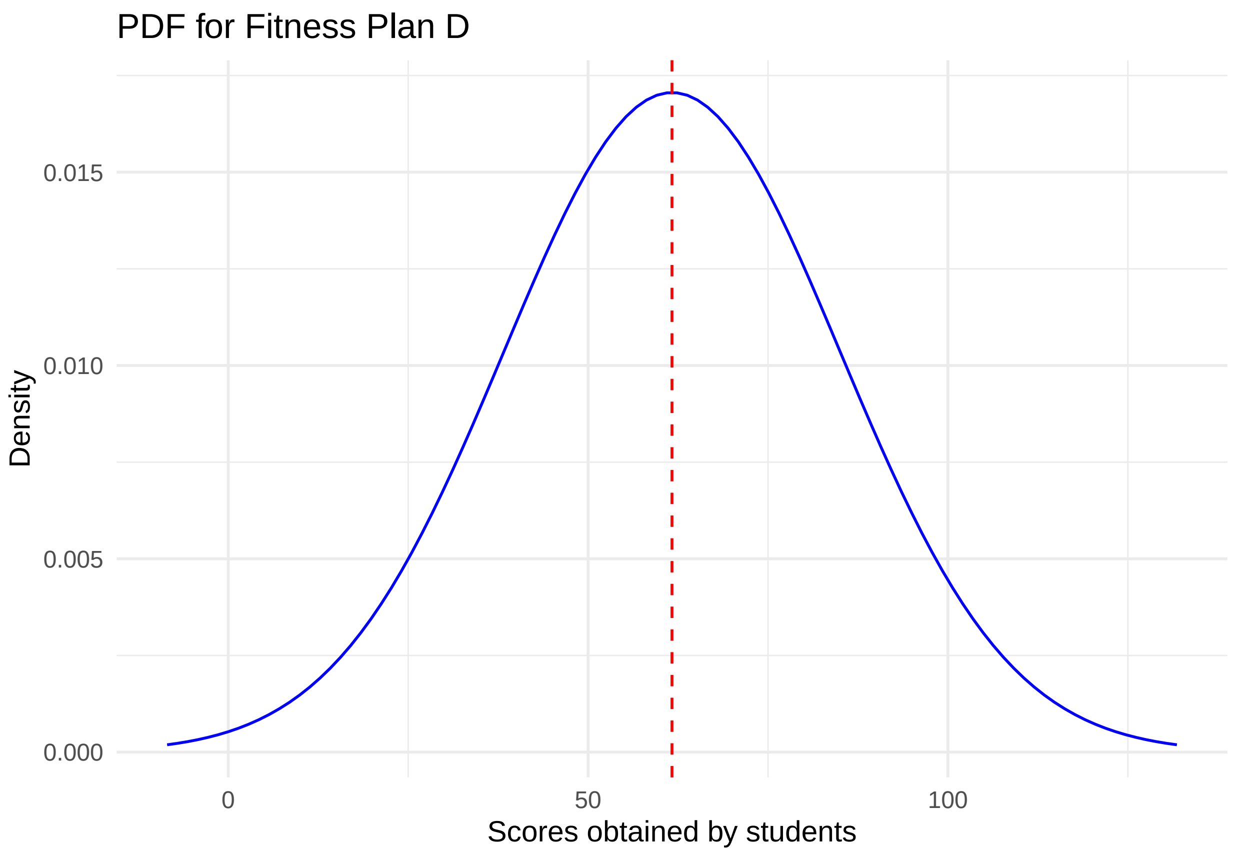
```r
# Function to create a PDF for a given column
create_bell_curve_plot <- function(column_data, column_name) {
  # Get the mean of the column
  mu <- mean(column_data)
  # Get the standard deviation of the column
  sigma <- sd(column_data)
  # Creating a sequence of values within a range around the mean (mu) of the data
  # with 100 evenly spread out data points.
  x <- seq(mu - 3*sigma, mu + 3*sigma, length.out = 100)
  # Using the dnorm function to calculate the probability density values
  # of a normal distribution at specific x-values.
  y <- dnorm(x, mean = mu, sd = sigma)
  ggplot() +
    geom_line(aes(x, y), color = "blue") +
    geom_vline(xintercept = mu, color = "red", linetype = "dashed") +
    labs(title = paste("PDF for Fitness Plan", column_name),
         x = "Scores obtained by students", y = "Density") +
    theme_minimal()
}

# Loop through all the fitness plans and create a PDF for them
for (col_name in colnames(data)) {
  plot <- create_bell_curve_plot(data[[col_name]], col_name)
  print(plot)
}
```


PDF for Fitness Plan A
Density
0.03
0.02
0.01
0.00
30
50
70
Scores obtained by students


PDF for Fitness Plan B
Density
0.020
0.015
0.010
0.005
0.000
0
25
50
75
100
Scores obtained by students

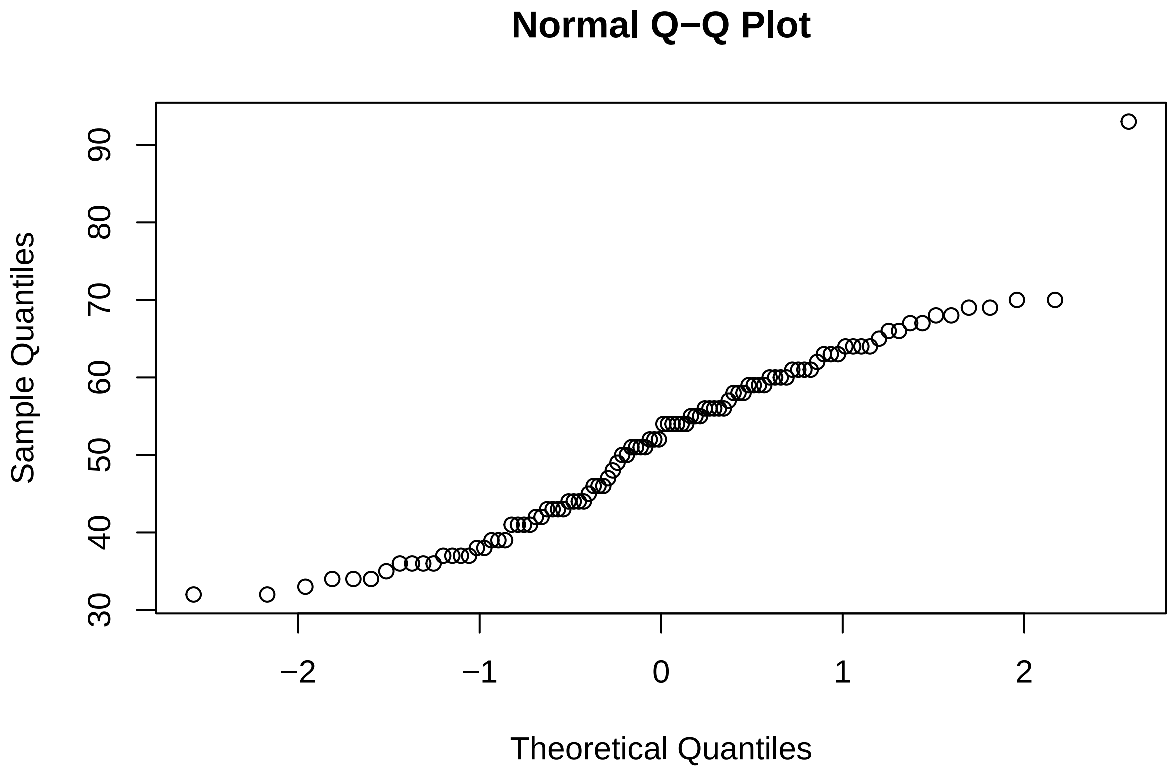

PDF for Fitness Plan C
Density
0.015
0.010
0.005
0.000
0
50
100
Scores obtained by students



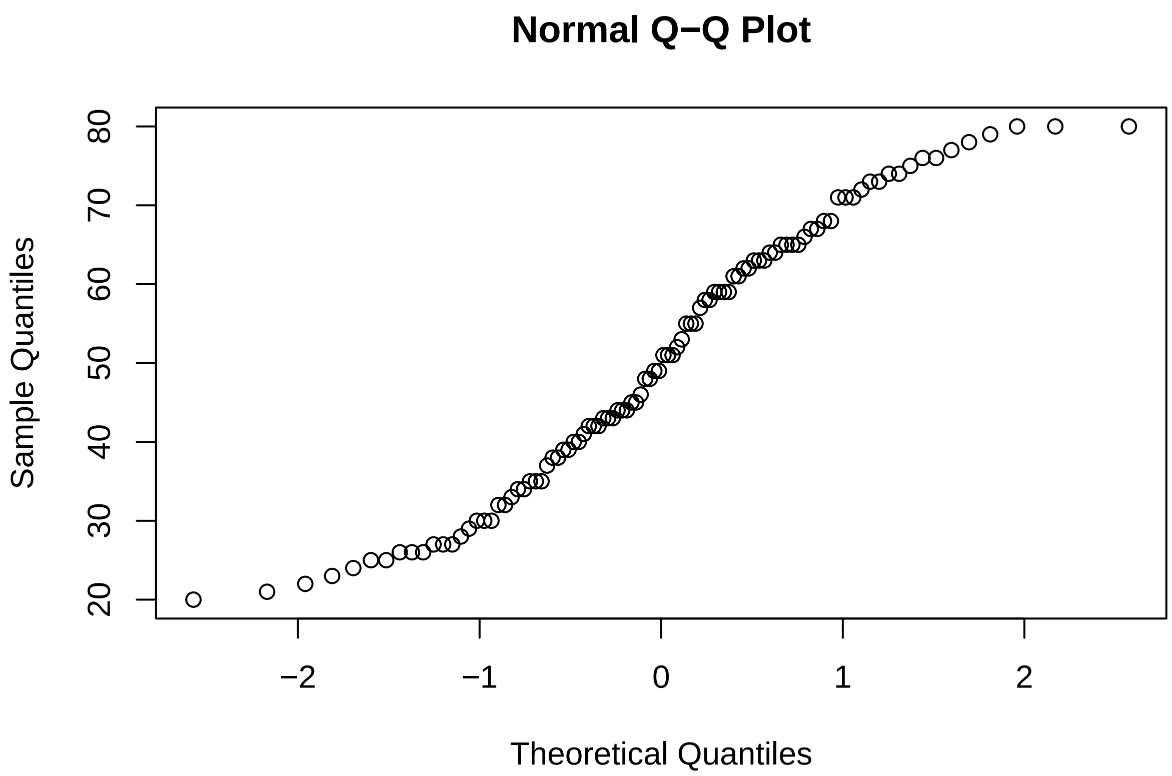
The plots justify the results of the pearson's moment coefficient of skewness.

We can also plot a Q-Q (quantile-quantile) plot of the data to visually assess whether the data points are approximately normally distributed.

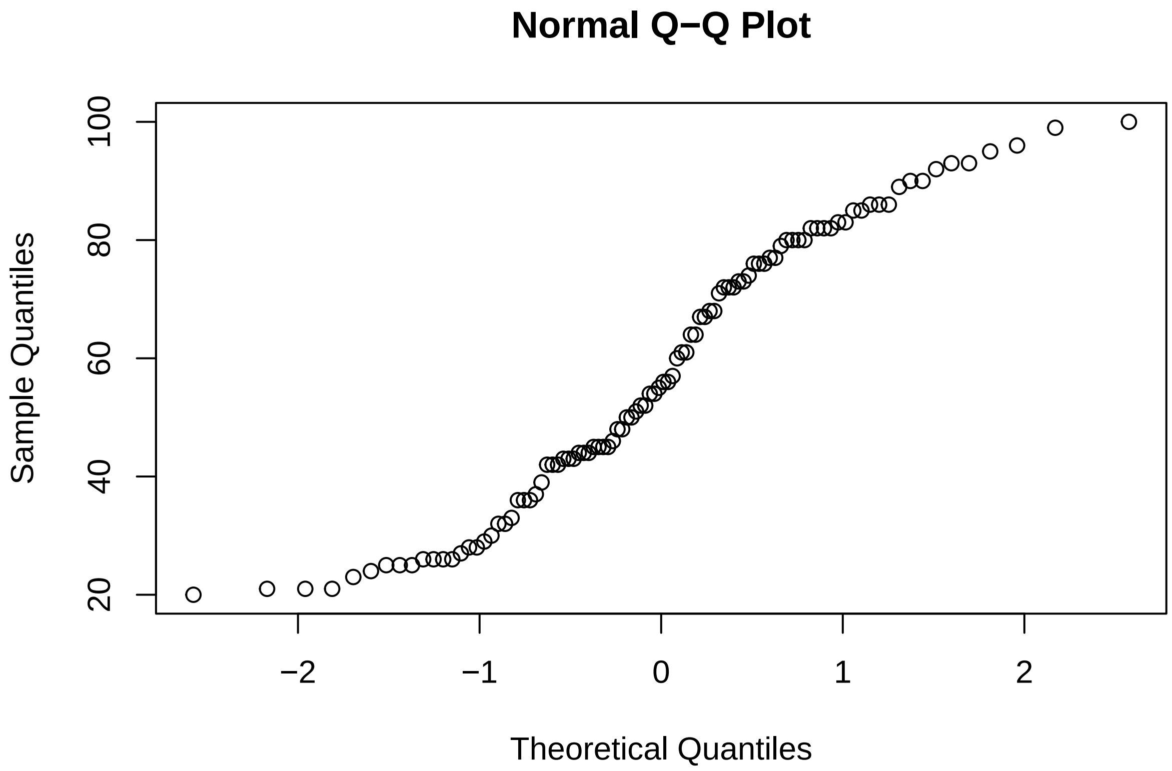
```
qqnorm(data$A)
```



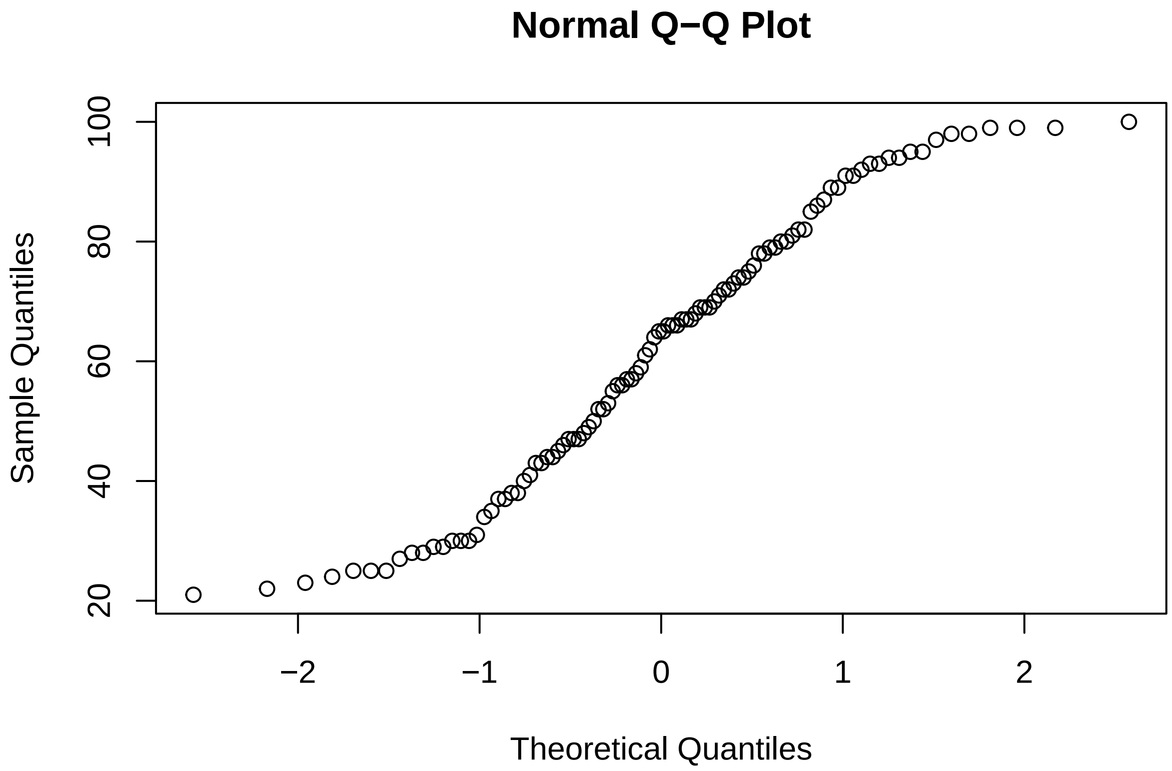
```r
qqnorm(data$B)
```



```
qqnorm(data$C)
```



```
qqnorm(data$D)
```



Because the PDF is approximately bell-shaped/normally distributed curve and the points in the Q-Q plot follow a roughly straight line, it suggests normality.

Problem 3

Is there any evidence to suggest a difference in the average marks obtained by students under different fitness plans? Explain what test are you using and why ? Define the hypothesis and the steps of testing. What does the output of this test signify ? (*Note: Assume the significance level to be 0.05*)

SOLUTION:

Step 1. Decide on the test.

As we know there are four different groups,we need to set up a one way ANOVA to demonstrate evidence to prove that one fitness plan is better than the other.

Step 2. Formulate Hypothesis.

To perform any tests, we first need to define the null and alternate hypothesis:

- Null Hypothesis: There is no significant difference among the groups.
- Alternate Hypothesis: There is a significant difference among the groups.

Step 3. Calculate SSW and SSB.

```r
data <- read.csv("students.csv", header = TRUE)

# Load required packages
if (!requireNamespace("readr", quietly = TRUE)) {
  install.packages("readr")
}
library(readr)

# Read CSV data
#data <- read_csv("your_file.csv")

# Calculate overall mean
overall_mean <- mean(as.matrix(data), na.rm = TRUE)

# Calculate column means
column_means <- colMeans(data, na.rm = TRUE)

# Calculate SSB
ssb <- sum((column_means - overall_mean)^2) * nrow(data)

# Calculate SSW
ssw <- sum((as.matrix(data) - rep(column_means, each = nrow(data)))^2)


cat("SSB: ", ssb, "\n")
```

```
## SSB:  8309.22
```

```
cat("SSW", ssw, "\n")
```

```
## SSW 150518.3
```

Step 4. Calculate MSB and MSW.

```
# Calculate degrees of freedom
n_total <- nrow(data)
n_columns <- ncol(data)
df_total <- n_total - 1
df_columns <- n_columns - 1
df_within <- n_total - n_columns

# Calculate MSB (Mean Square Between)
msb <- ssb / df_columns

# Calculate MSW (Mean Square Within)
msw <- ssw / df_within

# Print MSB and MSW
cat("MSB: ", msb, "\n")
```

```
## MSB:  2769.74
```

```
cat("MSW: ", msw, "\n")
```

```
## MSW:  1567.899
```

Step 5. Calculate the F-statistic.

The F-statistic for an ANOVA analysis is calculated using the formula:

$$F = \frac{SSB/(k-1)}{SSW/(n-k)} = \frac{MSB}{MSW}$$

Where:

- F is the F-statistic,

- SSB is the Between-Groups Sum of Squares,

- SSW is the Within-Groups Sum of Squares,

- k is the number of groups,

- n is the total number of observations.

- MSB is the Mean Square between Variation,

- MSW is the mean Square of variation Within the Group,

This formula is used to test for the presence of significant differences in means among the groups.

```
# Calculate the F-statistic
f_statistic <- msb / msw

# Print the F-statistic
cat("F-statistic: ", f_statistic, "\n")
```

```
## F-statistic:  1.766529
```

Step 6. Calculate p-value

```
# Calculate the degrees of freedom for the F-distribution
df1 <- df_columns
df2 <- df_within

# Calculate the p-value
# lower.tail is False because we are doing a right tailed test
p_value <- pf(f_statistic, df1, df2, lower.tail = FALSE)

# Print the p-value
cat("p-value: ", p_value, "\n")
```

```
## p-value:  0.1587336
```

Since the p value > 0.05 we do not have enough evidence to reject the null hypothesis. Hence there is no relation between the fitness plan the student is following and the marks obtained by the student.

Two-way ANOVA

About the Dataset

A community of pet lovers and trainers gathered for an exciting pet training event. With a total of 48 pets participating, each pet was given a Task (A/B/C/D) and a treat (I,II,III) for finishing it. The response times for each pet was recorded. All pets were assigned only one task and one treat.

The dataset can be found here

Problem 4

Which specific task exhibits the lowest average training time? Does the combination of different treats and tasks significantly influence the training time for pets?

```r
library(ggplot2)
library(readr)
library(stats)
```

```r
  data <- read.csv("pet_training.csv")
```

```r
anova_result <- aov(ResponseTime ~ Treat + Task + Treat:Task, data = data)
summary(anova_result)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## Treat        2 0.9374  0.4687  55.850 9.22e-12 ***
## Task         3 0.5835  0.1945  23.176 1.57e-08 ***
## Treat:Task   6 0.0526  0.0088   1.044    0.413    
## Residuals   36 0.3021  0.0084                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The combination of different treats and tasks significantly influences the training time for pets. This conclusion is based on the p-values associated with the main effects of "Treat" and "Task" in the ANOVA summary. The main effect of "Treat" has a highly significant p-value ($p < 0.001$) indicating that different treats have a significant influence on training time. The main effect of "Task" also has a highly significant p-value ($p < 0.001$) indicating that different tasks have a significant influence on training time.

As seen in the above output, "Treat:Task" is not statistically significant ($p = 0.413$).

```r
library(dplyr)
# Calculating average training time for each task
avg_training_time <- data %>%
  group_by(Task) %>%
  summarize(AvgTrainingTime = mean(ResponseTime))

# Identifying the task with the least average training time
task_with_least_time <- avg_training_time %>% filter(AvgTrainingTime == min(AvgTrainingTime))

print(task_with_least_time)
```

```
## # A tibble: 1 x 2
##   Task  AvgTrainingTime
##   <chr>           <dbl>
## 1 B               0.241
```

The task that requires the least training time is Task B with an average training time of approximately 0.2410 (in tens of minutes).

Problem 5

Does the choice of treats significantly impact the training time for different tasks? Which specific combinations of treats and tasks lead to the most significant differences in training time? (*Note: Assume the significance level to be 0.05*)

```r
library(readr)
library(dplyr)
library(multcomp)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
## Loading required package: TH.data

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

```r
library(stats)
```

```r
data$Treat <- factor(data$Treat)

# Perform two-way ANOVA
anova_result <- aov(ResponseTime ~ Treat * Task, data = data)

# Tukey's post hoc analysis
posthoc_treat <- glht(anova_result, linfct = mcp(Treat = "Tukey"))
```

```
## Warning in mcp2matrix(model, linfct = linfct): covariate interactions found --
## default contrast might be inappropriate
```

```r
posthoc_summary_treat <- summary(posthoc_treat)
print(posthoc_summary_treat)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = ResponseTime ~ Treat * Task, data = data)
##
## Linear Hypotheses:
##                 Estimate Std. Error t value Pr(>|t|)
## II - I == 0      0.11458    0.06478   1.769   0.1944
## III - I == 0     0.28979    0.06478   4.474   <0.001 ***
## III - II == 0    0.17521    0.06478   2.705   0.0274 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```r
# Summarize the ANOVA results
anova_summary <- summary(anova_result)

print(anova_summary)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Treat         2 0.9374  0.4687  55.850 9.22e-12 ***
## Task          3 0.5835  0.1945  23.176 1.57e-08 ***
## Treat:Task   6 0.0526  0.0088   1.044    0.413
## Residuals   36 0.3021  0.0084
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:

Question: Does the choice of treats significantly impact the training time for different tasks?

Answer: Yes, the choice of treats significantly impacts the training time for pets. This conclusion is based on the highly significant p-value ($p < 0.001$) associated with the "Treat" factor in the ANOVA summary. The significant p-value indicates that different treats have a significant influence on the training time for pets.

Question: Which specific combinations of treats and tasks lead to the most significant differences in training time?

Answer: The post hoc tests for the "Treat" factor(using a significance threshold of 0.05) did not identify any specific combinations of treats that exhibit statistically significant differences in training time. None of the pairwise comparisons including "II - I," "III - I," and "III - II" show statistically significant differences based on the adjusted p-values. This suggests that, in this particular analysis and dataset, there are no specific combinations of treats and tasks that lead to statistically significant differences in training time.