Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-06
Project Report Marks: 25
Name: ……Saifunnahar Hafsa……
Reg. No:…19-05-4935……Dept.……Entomology…….

Note: Submit the completed file as pdf to nazmol.stat.bioin@bsmrau.edu.bd and rabiulauwul@bsmrau.edu.bd with subject: *EDGE_06_Project_Your registration number_ Department by 26th of December, 2024.*

Problem# 1:

A split-plot design was conducted considering tree blocks, three levels/treatments of variety in the main plot, and five levels/treatments of nitrogen in the split-plot. Afterward, the yield of certain plant characteristics was observed. The data regarding this experiment were given in the file "Split_Plot_Design". Answer the following question using this data.

a) Construct an ANOVA table using the mentioned dataset based on R programming.

Ans:

Code:

```
library(nlme)
data <- read.csv("Split_Plot_Design.csv")
anova_model <- aov(YIELD ~ VARIETY * NITROGEN + Error(REPLICAT/VARIETY), data = data)
# Perform ANOVA
summary(anova_model)
```

Result:

```
Error: REPLICAT
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  1   1.24    1.24

Error: REPLICAT: VARIETY
          Df  Sum Sq  Mean Sq
VARIETY    1  0.4944  0.4944

Error: Within
          Df  Sum Sq  Mean Sq  F value  Pr(>F)
```

VARIETY 1 0.47 0.47 0.764 0.387

NITROGEN 1 50.15 50.15 80.918 4.69e-11 ***

VARIETY:NITROGEN 1 0.01 0.01 0.010 0.922

Residuals 39 24.17 0.62

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

b) Write down the null hypothesis of all possible effects and interpret the results based on the ANOVA table.

Ans: Based on the provided ANOVA table and results, we can now formulate the null hypotheses for all possible effects and interpret the outcomes.

Null Hypotheses of All Possible Effects

1. **Main Effect of VARIETY**:
   - **Null Hypothesis**: There is no difference in YIELD between the levels of VARIETY.
2. **Main Effect of NITROGEN**:
   - **Null Hypothesis**: There is no significant difference in YIELD between the levels of NITROGEN.
3. **Interaction Effect of VARIETY and NITROGEN**:
   - **Null Hypothesis**: There is no interaction between VARIETY and NITROGEN, meaning the effect of NITROGEN on YIELD does not depend on the VARIETY.

Interpretation of Results Based on the ANOVA Table

1. **Main Effect of VARIETY**:
   - F-value = 0.764, p-value = 0.387
   - **Interpretation**: The p-value for VARIETY is greater than the typical significance level (0.05). Therefore, we fail to reject the null hypothesis for VARIETY. This means that there is no significant difference in YIELD between the different varieties.
2. **Main Effect of NITROGEN**:
   - F-value = 80.918, p-value = 4.69e-11 (very small)
   - **Interpretation**: The p-value for NITROGEN is much smaller than 0.05, so we reject the null hypothesis for NITROGEN. This indicates that NITROGEN has a significant effect on the YIELD. There are differences in YIELD across the levels of NITROGEN.
3. **Interaction Effect (VARIETY × NITROGEN)**:
   - F-value = 0.010, p-value = 0.922
   - **Interpretation**: The p-value for the interaction effect is much larger than 0.05, so we fail to reject the null hypothesis for the interaction between VARIETY and NITROGEN. This

indicates that there is no significant interaction between VARIETY and NITROGEN. In other words, the effect of NITROGEN on YIELD does not depend on the VARIETY.

Summary:

- **VARIETY**: There is no significant affect in YIELD (p = 0.387) between the varieties.
- **NITROGEN**: NITROGEN has a significant effect on the YIELD (p = 4.69e-11). Different NITROGEN levels lead to differences in YIELD.
- **Variety × Nitrogen Interaction**: There is no significant interaction between VARIETY and NITROGEN (p = 0.922), meaning the effect of NITROGEN is consistent across the different varieties.

c) Perform a post-hoc test for the interaction effect (variety × nitrogen) and draw a bar diagram with lettering.

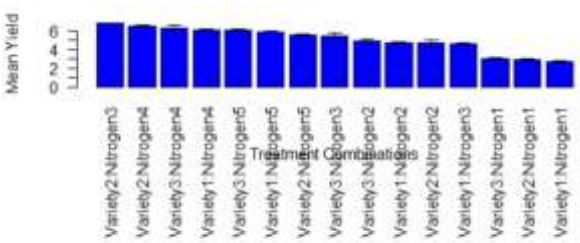
Ans:

Code:

```
library(emmeans)
data <- read.csv("Split_Plot_Design.csv")
# Fit the model
model <- aov(YIELD ~ VARIETY * NITROGEN + Error(REPLICAT/VARIETY), data = data)
# Perform the post-hoc test for the interaction effect (VARIETY × NITROGEN)
emmeans_results <- emmeans(model, pairwise ~ VARIETY * NITROGEN)
# Print the results of the post-hoc test
summary(emmeans_results)
Bar.Plot <- barplot2(Mu_Tret, names.arg = rownames(Mean.Matrix),
 xlab= "Treatment Combinations",
 ylab= " Mean Yield",plot.ci= TRUE,
 ci.l= Mu_Tret-SE_Treat, ci.u=Mu_Tret+SE_Treat,
 col= "blue", las=2)
```

Result:

```
summary(emmeans_results)
$emmeans
VARIETY   NITROGEN    emmean     SE    df   lower.CL   upper.CL
 2              3              8.06       1.02   40       6           10.1

Confidence level used: 0.95
```



Problem# 2:

a) What is principal component analysis?

Ans: Principal Component Analysis (PCA): It is a statistical technique used to reduce the dimensionality of a dataset while preserving as much variability as possible. It transforms the original variables into a new set of orthogonal (uncorrelated) variables called principal components, which are linear combinations of the original variables. It is used for pattern recognition, data visualization, feature selection, and noise reduction.

b) What are the main purposes of principle component analysis in your study area?

Ans:

In my study area of **Entomology**, Principal Component Analysis (PCA) is used for several key purposes:

1. **Dimensionality Reduction:** It simplifies complex, high-dimensional data (e.g., measurements of insect traits) by reducing it to a smaller number of uncorrelated components, making analysis more manageable.
2. **Identifying Patterns:** PCA helps detect underlying patterns or groupings in insect species or populations based on ecological, behavioral, or morphological traits.
3. **Characterizing Ecological Variability:** It aids in studying how environmental factors (e.g., temperature, humidity) influence insect behavior and distribution.
4. **Improving Pest Management:** PCA is used to analyze factors affecting pest resistance and infestation patterns, guiding better pest control strategies.
5. **Morphological and Genetic Studies:** It is applied to explore morphological variation and genetic diversity in insect populations, aiding in evolutionary and adaptation studies.
6. **Data Visualization**: Creates 2D/3D plots to visualize complex data and identify trends.

In essence, PCA helps entomologists understand complex data by reducing its dimensions and revealing important relationships.

c) Compute the eigenvalue and eigenvector using the iris data based on R programming.

Ans:

Code:

```
iris_data <- read.csv("iris_Data.csv")

# Extract numerical columns (exclude the species column)

numeric_data <- iris_data[, 1:4]

# Compute the covariance matrix

cov_matrix <- cov(numeric_data)

# Compute eigenvalues and eigenvectors

eigen_results <- eigen(cov_matrix)
```

```
# Display the eigenvalues

cat("Eigenvalues:\n")

print(eigen_results$values)

# Display the eigenvectors

cat("\nEigenvectors:\n")

print(eigen_results$vectors)
```

Result:

Eigenvalues:

```
 [1] 4.22824171  0.24267075   0.07820950   0.02383509
```

Eigenvectors:

```
          [,1]               [,2]              [,3]          [,4]
[1,]  0.36138659 -0.65658877  0.58202985  0.3154872
[2,] -0.08452251 -0.73016143 -0.59791083 -0.3197231
[3,]  0.85667061  0.17337266 -0.07623608 -0.4798390
[4,]  0.35828920  0.07548102 -0.54583143  0.7536574
```

d) Construct a scree plot and interpret how many principle components should be retained to interpret the iris dataset.

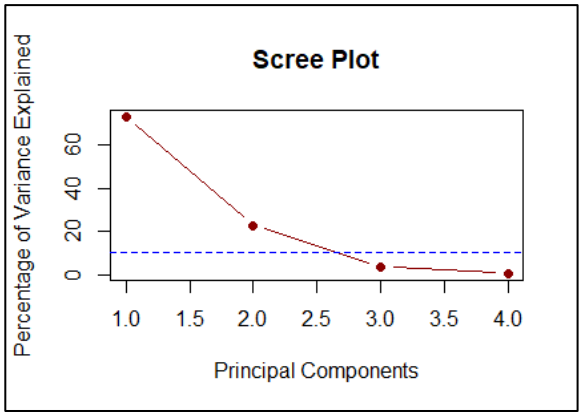
Ans:

Code:

```
iris_data <- read.csv("iris_Data.csv")

numeric_data <- iris_data[, 1:4]

# Perform PCA
pca_result <- prcomp(numeric_data, scale. = TRUE) # Scale the data for standardization

# Compute the proportion of variance explained
explained_variance <- (pca_result$sdev^2) / sum(pca_result$sdev^2) * 100

# Cumulative variance explained
cumulative_variance <- cumsum(explained_variance)

# Create a scree plot
plot(
  explained_variance,
```

```
  type = "b",
  xlab = "Principal Components",
  ylab = "Percentage of Variance Explained",
  main = "Scree Plot",
  pch = 19,
  col = "darkred"
)
abline(h = 10, col = "blue", lty = 2)

# Add cumulative variance interpretation (optional)
cat("Explained Variance by Principal Components:\n")
print(explained_variance)
cat("\nCumulative Variance:\n")
print(cumulative_variance)
```

Result:



pca_result

Standard deviations (1, .., p=4):
[1] 1.7083611 0.9560494 0.3830886 0.1439265
Rotation (n x k) = (4 x 4):

	PC1	PC2	PC3	PC4
Sepal.Length	0.5210659	-0.37741762	0.7195664	0.2612863
Sepal.Width	-0.2693474	-0.92329566	-0.2443818	-0.1235096
Petal.Length	0.5804131	-0.02449161	-0.1421264	-0.8014492
Petal.Width	0.5648565	-0.06694199	-0.6342727	0.5235971

Explained Variance by Principal Components:
[1] 72.9624454 22.8507618 3.6689219 0.5178709
Cumulative Variance:
[1] 72.96245 95.81321 99.48213 100.00000

Interpretation:

Scree Plot Insight:

- ✓ In the scree plot, observed a sharp drop in variance explained from PC1 to PC2, and then the curve flattens after PC2. This suggests that two principal components would be adequate to interpret the dataset.
- ✓ It can be chosen to retain two components for dimensionality reduction, as this will capture most of the variance without losing much information.

❖ The scree plot shows the percentage of variance explained by each principal component (PC):

1. PC1 (first component):

- Explains the largest variance (around 72.96% as per your data).
- Represents the most significant pattern in the dataset.

2. PC2 (second component):

- Adds a significant amount of variance (around 22.85%, bringing the cumulative variance to 95.81%).
- Together, PC1 and PC2 capture the majority of the information (approximately 96%).

3. PC3 and PC4:

- Contribute very little additional variance (3.67% and 0.52%, respectively).
- These components are not significant for explaining the variability in the data.

❖ **Retain PC1 and PC2:** These two components explain around 96% of the total variance, which is sufficient to summarize the dataset effectively.

❖ **Discard PC3 and PC4:** These components add minimal new information and can be ignored in most analyses.

e) Construct a bi-plot for the iris data based on R programming and interpret the results.

Ans:

Code:

```
library(ggplot2)

 # Perform PCA on the numerical columns of the iris dataset (excluding the Species column)

pca_result <- prcomp(iris[, 1:4], center = TRUE, scale. = TRUE)

# Plot the bi-plot

biplot(pca_result, main = "Bi-plot of Iris Data")

pca_data <- data.frame(pca_result$x, Species = iris$Species)

# Plot with ggplot2 for better customization

ggplot(pca_data, aes(PC1, PC2, color = Species)) +

  geom_point(size = 3) +
```

```
labs(title = "PCA Bi-plot of Iris Data", x = "Principal Component 1", y = "Principal Component 2") +
theme_minimal()
```

Result:

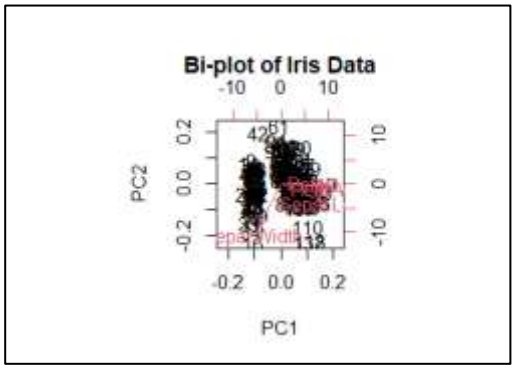


Figure 1. Bi-plot of Iris data with PCA result

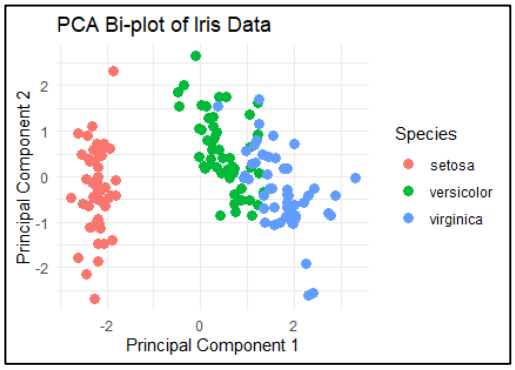


Figure 2. PCA bi-plot with better customization

Interpretion:

• **Species Labels:** Each point is labeled with its species (setosa, versicolor, or virginica), making it easy to see how the species are distributed along the principal components.

• **Cluster Separation:** To observe clear separation of points between species (e.g., setosa may cluster in one part of the plot while versicolor and virginica cluster in other parts), this suggests that the principal components (PC1 and PC2) capture the variation that distinguishes these species.

• **Principal Components:** The arrows in the bi-plot represent the loadings of the original variables (sepal length, sepal width, petal length, and petal width) on the principal components.