Design of Experiments, One-way ANOVA

Examples in R

Statistical Cross-disciplinary Collaboration and Consulting Lab (SC3L)

Oluwafunmibi Omotayo Fasanya

March 12, 2025

About SC3L

Hours of Operation

Monday: 11:00 AM - 1:00 PM, 3:00 PM - 5:00 PM

Tuesday: 12:00 PM - 4:00 PM

Wednesday: 11:00 AM - 3:00 PM

Thursday: 11:00 AM - 5:00 PM

Friday: 1:00 PM - 2:00 PM

What We Do

The Statistical Cross-disciplinary Collaboration and Consulting Lab (SC3L) is a free service available to students, faculty, and staff at the University of Nebraska who are in need of assistance with:

- MS thesis or doctoral dissertation
- Faculty research
- Statistical analysis support

Need Statistical Help?

1. Fill out the Google form on our website
   (required for record-keeping and consultant matching)
2. Schedule a meeting with an appropriate consultant

https://statistics.unl.edu/sc3lhelp-desk/

Outline

- Fundamentals of Experiment Design
  - What is an Experiment?
  - What is experimental design?
  - Principles of Experiment Design
  - Structure of Experiment Design
- One-Way ANOVA (Completely Randomized design)
  - Concept and Assumptions
- One-Way ANOVA (Completely Randomized design)
  - Hands-on Example in R

Fundamentals of Experimental Design

What is an Experiment? An experiment is the process of applying a treatment to experimental material or units and recording observations to answer a specific research question.

Treatment An experimental condition that is applied to the experimental material or unit.

Experimental Unit A subdivision of the experimental material such that different units can receive different treatments.

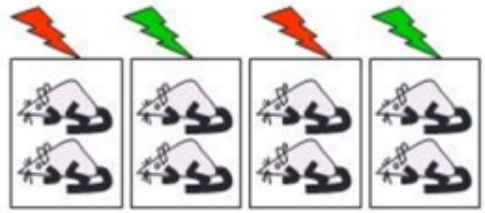
Experimental Unit

Scenario 1: Individual-Level Treatment

- A researcher is interested in evaluating the effects of two antibiotics on weight gain in mice.
- Treatment applied to individual mice (random injections).
- Experimental Unit: Individual mice.
- Total Experimental Units: 8.

Scenario 2: Group-Level Treatment

- The researcher now investigates the effects of two diets on weight gain.
- Treatment applied at the cage level (4 cages, 2 mice each).
- Experimental Unit: Cage.
- Total Experimental Units: 4.



Key Takeaway: The experimental unit is the smallest unit to which a treatment is independently applied.

What is Experimental Design?

- The method and procedure of planning experiments to obtain information about a specific question in an unbiased, clear, and precise manner.
- A research method that allows us to determine the effect of a single or multiple factors (treatments) on an outcome.
- The main goal of designing an experiment is to reduce error for a particular investigation.

Principles of Experimental Design

1. Randomization

- The process of randomly assigning treatments to experimental units.
- Ensures every treatment has an equal chance of being assigned to different experimental units.
- Helps eliminate bias from researchers' judgment.
- The most reliable method for creating homogeneous treatment groups without potential biases.

Example: Randomization

Example: Randomizing Treatment Assignment

- Suppose we have 20 experimental units and want to test two fertilizers (A and B).
- Instead of assigning fertilizers based on pre-existing conditions, we use randomization:
  - Randomly assign 10 units to Fertilizer A.
  - Randomly assign 10 units to Fertilizer B.
- This ensures each unit has an equal chance of receiving either treatment, reducing selection bias.

Principles of Experimental Design

2. Replication (Ensuring Reliability)

- Experiments should be repeated more than once to reduce variability in results.
- Each treatment is applied to multiple experimental units rather than just one.
- Replication helps reduce variability, increases significance, and improves confidence in conclusions.

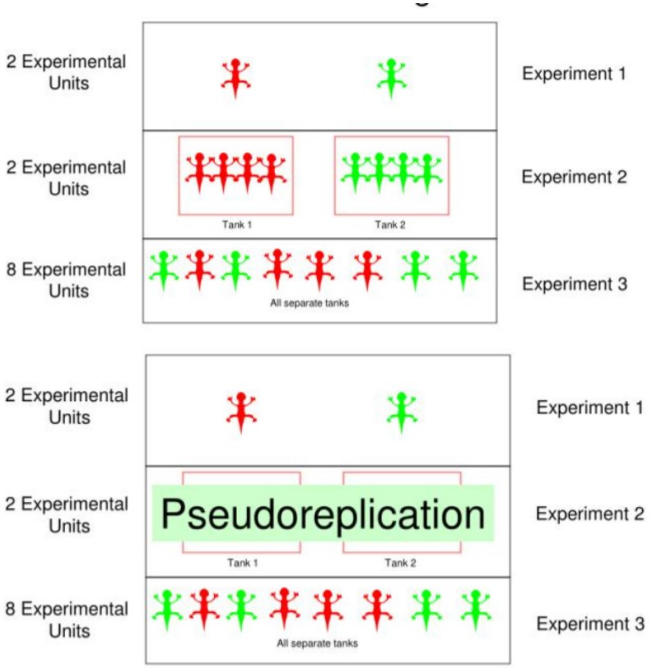
Example: Replication

Example: Examining Maize Varieties

- Suppose we want to examine the effect of four maize varieties. We divide a field into four parts and grow one variety in each.
- Applying replication: Instead of four large sections, we divide the field into 12 smaller plots and randomly assign:
  - Three plots to Variety A
  - Three plots to Variety B
  - Three plots to Variety C
  - Three plots to Variety D
- This reduces variability and improves the reliability of results.
- Note that we randomly assign each variety to multiple plots to reduce bias caused by soil fertility, moisture, or other environmental factors.
- Each variety is grown in several plots (not just one), ensuring that we obtain multiple independent observations for each variety.

Pseudoreplication

- Pseudoreplication occurs in an experiments where either treatments are not replicated or replicates are not statistically independent.
- True replication allows the estimation of variability between treatments.



- Pseudoreplication can lead to misleading conclusions due to confounding factors.
- If an external factor affects only one group, it can create false differences.
- The observed difference may not be due to the treatment itself but rather to an uncontrolled variable.
- Without proper replication, it's easy to mistake random chance for a real effect.

Principles of Experimental Design

3. Blocking (Reducing Variability)

- Blocking involves grouping similar subjects together before randomly assigning them to treatment groups.
- This helps reduce variability within treatment groups by ensuring they are balanced on key characteristics (blocking by age, then randomly assigning within each age group).

Example: Blocking

Example: Accounting for Soil Differences

- Suppose we are testing different fertilizers on crop yield, but the field has variations in soil quality.
- To account for this, we divide the field into blocks based on soil type (e.g., sandy, loamy, clay).
- Within each block, we randomly assign fertilizer treatments, ensuring soil differences do not confound results.

Experimental Design

Experimental design can be broken into:

- **Treatment Structure:** How treatments are arranged.
  - **Single Factor:** The treatment consists of a single variable factor.
  - **Multifactor:** Multiple factors are considered to analyze their combined effects.
    - Full Factorial
    - Fractional Factorial
    - Factorial with Control
    - Nested Structure
- **Design Structure:** How experimental units are grouped into homogeneous groups or blocks.
  - Completely Randomized Design:
    - Each experimental units is assigned to a treatment at random without accounting for any individual characteristics.
  - Randomized Block Design
  - Row-Column Design
  - Split-Plot Design
  - Criss-Cross Design

One-Way ANOVA

(Completely Randomized Design)

Concepts and Assumptions

What is One-Way ANOVA?

- A statistical technique to test for statistically significant differences between means of three or more independent groups
- Analyzes the variance among values
- "One-Way" refers to having only a single predictor/explanatory variable 'X' and one response variable 'Y'

Requirements for One-Way ANOVA

- **Dependent Variable**: Continuous
- **Independent Variable**: Categorical, consisting of three or more groups
- **Independence of Observations**: Samples or groups being compared must be independent, this implies
  - Participants in one group should not be part of another group
  - Individuals in one group must not influence those in another
  - If subjects were matched, or samples represent before/during/after measurements, use a different ANOVA type
- **Random Sampling**: Samples randomly selected from or representative of larger populations

Requirements for One-Way ANOVA

- **Normality Assumption**: Dependent variable should follow approximately normal distribution.
  - With larger sample sizes, minor violations of normality may still yield reliable results
- **Homogeneity of Variances:**
  - Variance of dependent variable should be roughly equal across all groups
- **No Extreme Outliers:**
  - The presence of significant outliers can distort results and should be examined before conducting the analysis.

Hypotheses for One-Way ANOVA

Null Hypothesis (H_0): No difference between the groups

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_t$$

Alternative Hypothesis (H_1): At least one group is different

$$H_1 : \text{at least one } \mu_i \text{ differs}$$

where μ_i is the population mean of the i-th group ($i = 1, 2, \ldots, t$).

One-Way ANOVA: Variation Components

Total Variation in the data is divided into:

- **Between-Group Variation**: Differences in means across groups.
- **Within-Group Variation**: Variability within each group due to individual differences.

The One-Way ANOVA Table:

	SS	**df**	**MS**	**F**	**p**
Between Groups	SSH	$t-1$	$MST = \frac{SSH}{t-1}$	$\frac{MSH}{MSE}$	F_{df_1, df_e}
Within Groups	SSE	$n-t$	$MSE = \frac{SSR}{n-t}$		
Total	SST	$n-1$			

F-Statistic in One-Way ANOVA

The test statistic for ANOVA is the F-statistic:

$$F = \frac{\text{Between-Group Variation}}{\text{Within-Group Variation}}$$

- A **larger F-value** suggests significant group differences.
- A **smaller F-value** suggests differences are due to random variation.

Data Set-Up for One-Way ANOVA

To perform a one-way ANOVA, your dataset should include:

- **At least two variables:**
  - **Independent Variable:** A categorical factor with at least three distinct groups.
  - **Dependent Variable:** A continuous variable.
- Each row in the dataset should correspond to a unique participant or experimental unit.
- Observations must be **independent**.

One-Way ANOVA

(Completely Randomized Design)

Hands-on Example in R

Experimental Design: Fertilizer Study

Objective: Evaluate the effect of three fertilizers on maize yield using a completely randomized design.

Fertilizer Treatments:

- **Nitrogen-Based** (Red)
- **Potassium-Enriched** (Blue)
- **Organic Compost** (Yellow)

Randomization and Replication:

- Each fertilizer is applied to **14 plots**.
- **Total plots:** 42
- Completely randomized design (CRD) layout.

Randomized Fertilizer Plot Layout

Organic compost		Potassium-enriched		Nitrogen-Based
Potassium-enriched		Nitrogen-Based		Potassium-enriched
Nitrogen-Based		Nitrogen-Based		Organic compost
Potassium-enriched		Organic compost		Organic compost
Organic compost		Potassium-enriched		Potassium-enriched
Nitrogen-Based		Organic compost		Nitrogen-Based
Potassium-enriched		Potassium-enriched		Organic compost
Organic compost		Organic compost		Nitrogen-Based
Nitrogen-Based		Nitrogen-Based		Potassium-enriched
Potassium-enriched		Organic compost		Nitrogen-Based
Potassium-enriched		Potassium-enriched		Organic compost
Organic compost		Nitrogen-Based		Potassium-enriched
Nitrogen-Based		Nitrogen-Based		Organic compost
Organic compost		Potassium-enriched		Nitrogen-Based

Figure: Completely Randomized Design(CRD) for Fertilizer Treatments

Dataset: Fertilizer and Maize Yield

Objective: Evaluate the impact of three different fertilizer treatments on maize yield.

Yield Data (kg per plot):

Treatment	Yield Values
TRT 1	30.1, 21.1, 29.3, 24.5, 29.1, 25.4, 31.8, 30.7, 26.6, 22.5, 26.0, 21.7, 34.8, 35.4
TRT 2	28.2, 22.3, 20.6, 16.7, 22.8, 22.7, 28.1, 28.3, 24.7, 19.2, 23.6, 19.0, 34.3, 31.3
TRT 3	24.4, 20.2, 23.7, 18.5, 27.7, 26.7, 30.6, 24.0, 26.0, 23.6, 15.5, 16.3, 23.4, 19.8

Hypothesis for Fertilizer and Maize Yield Study

Null Hypothesis (H_0): There is no significant difference in maize yield among the three fertilizer treatments.

$$H_0 : \mu_{\text{Fertilizer 1}} = \mu_{\text{Fertilizer 2}} = \mu_{\text{Fertilizer 3}}$$

Alternative Hypothesis (H_1): At least one fertilizer group shows a significant difference in yield

$$H_1 : \text{At least one } \mu_i \text{ differs}$$

Step 1: Load Required Libraries

```r
# Load required libraries
library(tidyr)     # Data wrangling (e.g., pivot_longer)
library(dplyr)     # Data manipulation and summarization
library(ggplot2)   # Data visualization
library(car)       # Assumption tests
```

Step 2: Load and Transform Data

```r
# Load dataset
dat_CRD <- data.frame(
  Sample = 1:14,
  TRT1 = c(30.1, 21.1, 29.3, 24.5, 29.1, 25.4, 31.8,
           30.7, 26.6, 22.5, 26.0, 21.7, 34.8, 35.4),
  TRT2 = c(28.2, 22.3, 20.6, 16.7, 22.8, 22.7, 28.1,
           28.3, 24.7, 19.2, 23.6, 19.0, 34.3, 31.3),
  TRT3 = c(24.4, 20.2, 23.7, 18.5, 27.7, 26.7, 30.6,
           24.0, 26.0, 23.6, 15.5, 16.3, 23.4, 19.8)
)

# Convert data to long format
dat_CRD_long <- pivot_longer(dat_CRD, cols = -Sample,
                             names_to = "Treatment",
                             values_to = "Response")
dat_CRD_long$Treatment <- as.factor(dat_CRD_long$Treatment)
```

Step 3: Data Visualization - Boxplot

Boxplot of Treatment Effect on Response

```
# Boxplot
ggplot(dat_CRD_long, aes(x = Treatment,y = Response,fill = Treatment)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Treatment Effect on Response",x = "Treatment", y = "Response")
```

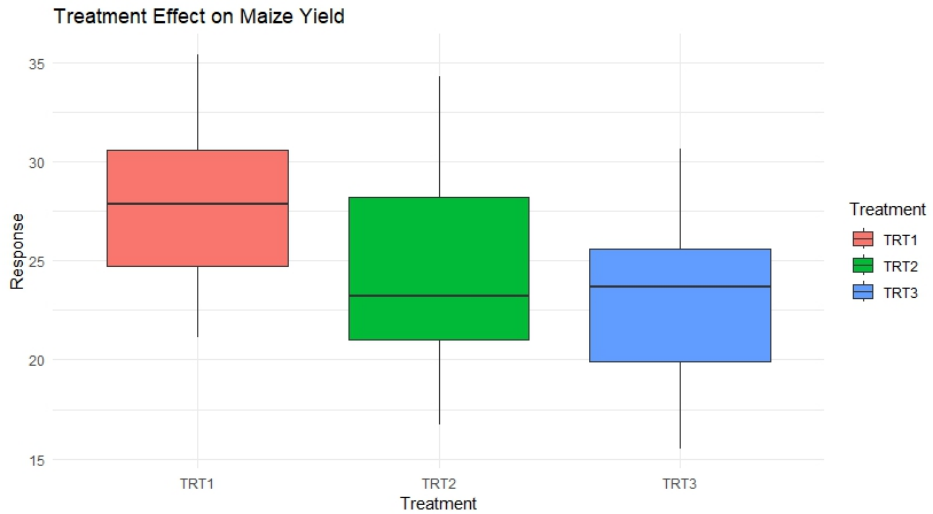


Figure: Boxplot of Treatment Effect on Response

Step 3: Data Visualization - Histogram

Histogram of Maize Yield Distribution

```
# Histogram
ggplot(dat_CRD_long, aes(x = Response)) +
  geom_histogram(aes(y = ..density..),bins = 15,fill = "skyblue",color = "black", alpha = 0.7) +
  geom_density(color = "red", size = 0.5) +
  theme_minimal() +
  labs(title = "Distribution of Maize Yield",x = "Response", y = "Density")
```

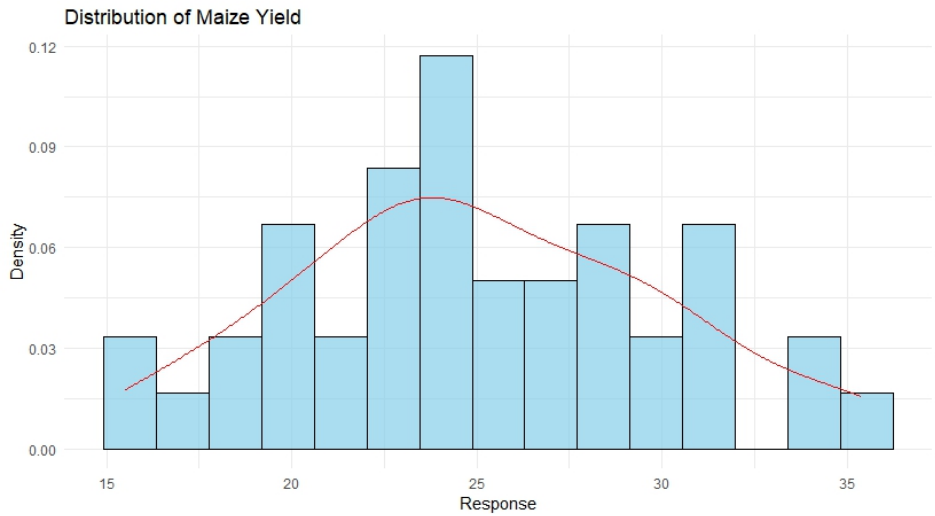


Figure: Histogram of Maize Yield Distribution

Step 5: Running One-Way ANOVA

ANOVA Model

```r
# Perform ANOVA
anova_mod <- aov(Response ~ Treatment, data = dat_CRD_long)
summary(anova_mod)
```

ANOVA Output:

```
                Df Sum Sq Mean Sq F value Pr(>F)
    Treatment    2  176.0   88.00   4.051  0.0252 *
    Residuals   39  847.2   21.72
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Step 5: Assumption Checks for ANOVA

Residual Analysis

```
# Plot residuals
plot(anova_mod)
```

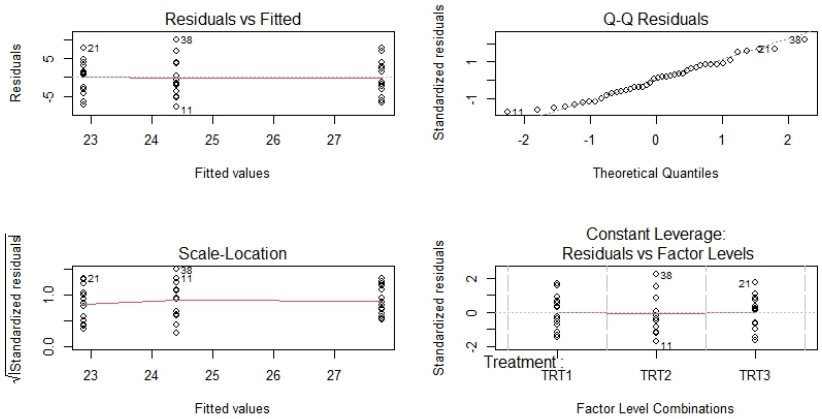


Figure: Residual Diagnostic Plot

Shapiro-Wilk Normality Test

```
# Shapiro-Wilk test on residuals
aov_residuals <- residuals(object = anova_mod)
shapiro.test(x = aov_residuals)
```

```
Shapiro-Wilk normality test
data:  aov_residuals
W = 0.97533, p-value = 0.4895
```

Levene's Test for Homogeneity of Variance

```
# Homogeneity of variance
leveneTest(aov_residuals ~ dat_CRD_long$Treatment)
```

```
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  2  0.2274 0.7976
      39
```

Step 6: Post-hoc Analysis

Tukey's HSD Test

```
tukey_test <- TukeyHSD(anova_mod)
print(tukey_test)
```

Post-hoc Result

```
Tukey multiple comparisons of means
  95% family-wise confidence level

Fit: aov(formula = Response ~ Treatment, data = dat_CRD_long)

$Treatment
                diff       lwr        upr     p adj
TRT2-TRT1 -3.371429 -7.663250  0.9203924 0.1482812
TRT3-TRT1 -4.900000 -9.191821 -0.6081790 0.0220928
TRT3-TRT2 -1.528571 -5.820392  2.7632496 0.6635876
```

Summary

- One-Way ANOVA helps determine if treatment groups differ significantly.
- Assumptions checked: normality and homogeneity of variance.
- Tukey's HSD used for post-hoc pairwise comparisons.

Effect of Feed Supplements on Pig Growth Rate

Feed supplementation is a common practice in commercial pig farming to enhance growth rates. This study assesses whether different feed supplements affect weight gain in weaned piglets.

Experimental Groups:

- **Group A**: Standard feed (control)
- **Group B**: Feed + Probiotic supplement
- **Group C**: Feed + Enzyme supplement
- **Group D**: Feed + Essential oil blend

Weight gain was measured over a 6-week period following weaning.

Weight Gain by Feed Group

Objective: Evaluate the impact of of Feed Supplements on Pig Growth Rate.

Pig ID	Feed Group	Weight Gain (kg)
P01	Standard	12.8
P02	Standard	13.5
P03	Standard	11.9
P04	Standard	12.3
P05	Standard	14.1
P06	Standard	13.0
P07	Probiotic	15.2
P08	Probiotic	16.4
P09	Probiotic	14.8
P10	Probiotic	15.9
P11	Probiotic	17.2
P12	Probiotic	16.1
P13	Enzyme	14.5
P14	Enzyme	15.3
P15	Enzyme	13.9
P16	Enzyme	14.8
P17	Enzyme	15.6
P18	Enzyme	14.2
P19	Essential Oil	13.7
P20	Essential Oil	14.4
P21	Essential Oil	12.9
P22	Essential Oil	13.5
P23	Essential Oil	14.8
P24	Essential Oil	13.2

Research Hypothesis for One-way ANOVA

Null Hypothesis (H_0): There is no significant difference in mean weight gain among the four feed supplement groups.

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$$

Alternative Hypothesis (H_1): At least one feed supplement group shows a significant difference in mean weight gain.

$$H_1 : \text{At least one } \mu_i \text{ differs from the others}$$

Step 2: Load the Data and check the structure

```r
# Creating a data frame
data <- data.frame(
  ID = c("P01", "P02", "P03", "P04", "P05", "P06",
         "P07", "P08", "P09", "P10", "P11", "P12",
         "P13", "P14", "P15", "P16", "P17", "P18",
         "P19", "P20", "P21", "P22", "P23", "P24"),
  Group = c(rep("Standard",6),rep("Probiotic",6), rep("Enzyme",6),rep("Essential Oil",6)),
  Weight = c(12.8, 13.5, 11.9, 12.3, 14.1, 13.0,
             15.2, 16.4, 14.8, 15.9, 17.2, 16.1,
             14.5, 15.3, 13.9, 14.8, 15.6, 14.2,
             13.7, 14.4, 12.9, 13.5, 14.8, 13.2)
)

# Convert Group to factor
data$Group <- factor(data$Group, levels = c("Standard", "Probiotic", "Enzyme", "Essential Oil"))

# Looking at the first few rows
head(data)

# Data Summary
summary(data)

# Data Structure
str(data)
```

Step 3: Data Visualization - Boxplot

Boxplot of Feed Supplement on Weight

```
# Boxplot
ggplot(dat, aes(x = Group, y = Weight, fill = Group)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Boxplot of Feed Supplement on Weight", x = "Group", y = "Weight")
```

Figure: Boxplot of Feed Supplement on Weight

Step 3: Data Visualization - Histogram

Histogram of Maize Yield Distribution

```
 # Histogram
ggplot(dat, aes(x = Weight)) +
  geom_histogram(aes(y = ..density..), bins = 15, fill = "skyblue", color = "black", alpha = 0.7) +
  geom_density(color = "red", size = 0.5) +
  theme_minimal() +
  labs(title = "Distribution of  Weight", x = "Response", y = "Density")
```

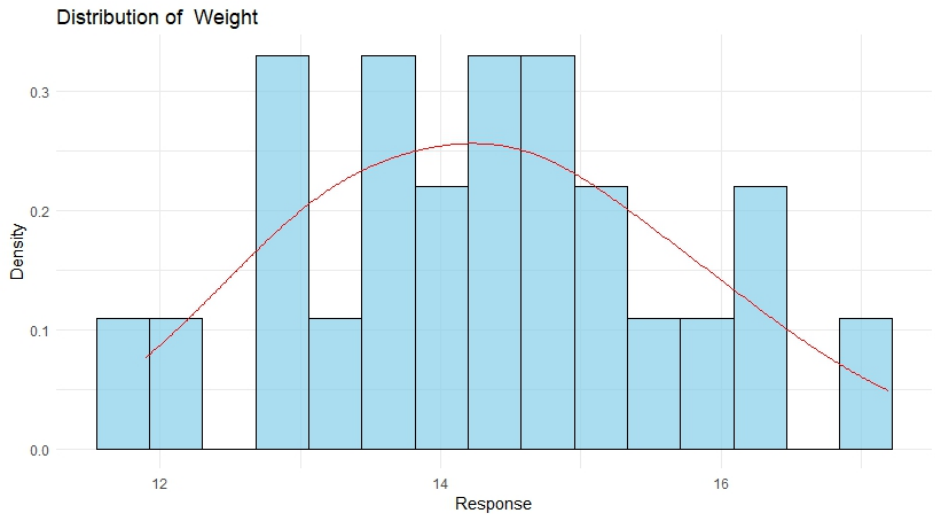


Figure: Histogram of Maize Yield Distribution

Step 5: Running One-Way ANOVA

ANOVA Model

```r
# Perform ANOVA
anova_mod2 <- aov(Weight ~ Group, data = dat)
summary(anova_mod2)
```

ANOVA Output:

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Group         3  30.04  10.014  17.31   8.83e-06 ***
Residuals    20  11.57   0.578
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Step 5: Assumption Checks for ANOVA

Residual Analysis

```
# Plot residuals
plot(anova_mod2)
```

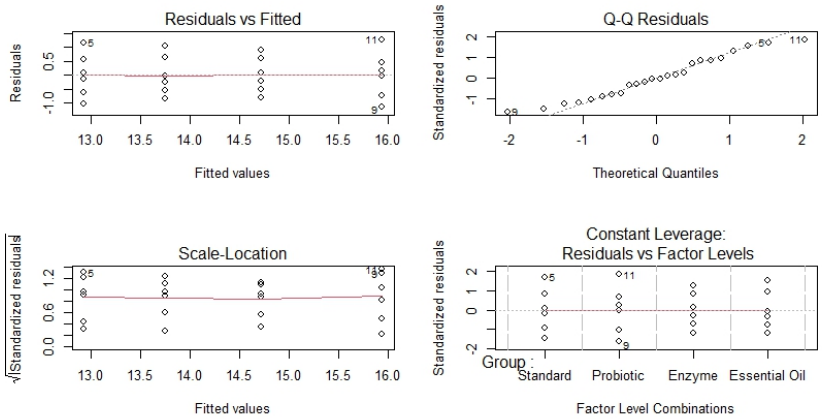


Figure: Residual Diagnostic Plot

Shapiro-Wilk Normality Test

```
# Shapiro-Wilk test on residuals
aov_residuals <- residuals(object = anova_mod2)
shapiro.test(x = aov_residuals)
```

```
Shapiro-Wilk normality test
data:  aov_residuals
W = 0.96241, p-value = 0.4887
```

Levene's Test for Homogeneity of Variance

```
# Homogeneity of variance
leveneTest(aov_residuals ~ dat$Group)  # Levene's test
```

```
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  3  0.0852 0.9673
      20
```

Step 6: Post-hoc Analysis

Tukey's HSD Test

```r
# Post-hoc tests (Tukey's HSD)
tukey_test <- TukeyHSD(anova_mod2)
print(tukey_test)
```

Post-hoc Result

```
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = Weight ~ Group, data = dat)

$Group
                              diff        lwr         upr     p adj
Probiotic-Standard       3.0000000  1.7709090  4.22909102 0.0000068
Enzyme-Standard          1.7833333  0.5542423  3.01242435 0.0031467
Essential Oil-Standard   0.8166667 -0.4124244  2.04575769 0.2762897
Enzyme-Probiotic        -1.2166667 -2.4457577  0.01242435 0.0529681
Essential Oil-Probiotic -2.1833333 -3.4124244 -0.95424231 0.0003949
Essential Oil-Enzyme    -0.9666667 -2.1957577  0.26242435 0.1570260
```

Questions?