DH 302 Spring 2025 Assignment 02

SAURAV KUMAR (21D070063)

February 18, 2025

Problem 01 [25 points]

1a. Null Hypothesis (H_0)

The null hypothesis states that there is no significant difference in the improvement of quality of life scores between the exercise group and the control group.

1b. Conclusion from the Dotplot

The dotplot suggests that the exercise group tends to show higher improvements than the control group; however, statistical testing is needed to confirm this observation.

1c. P-Value Interpretation

A p-value of 0.00866 indicates that, if the null hypothesis were true, there is only a 0.87% chance of obtaining results as extreme as those observed. This low p-value provides evidence against the null hypothesis.

1d. Conclusion for $\alpha = 0.01$

Since the p-value (0.00866) is less than $\alpha = 0.01$, we reject the null hypothesis. This suggests that the exercise program leads to a statistically significant improvement in quality of life scores.

1e. P-Value for the Nondirectional Test

For a two-tailed test, the p-value is:

$$2 \times 0.00866 = 0.01732.$$

1f. Conclusion for the Nondirectional Test

Since $0.01732 > 0.01$, we fail to reject the null hypothesis in a two-tailed test at the 1% significance level.

Problem 02 [10 points]

2a. Boxplot, Violin, and Ridgeline Plot

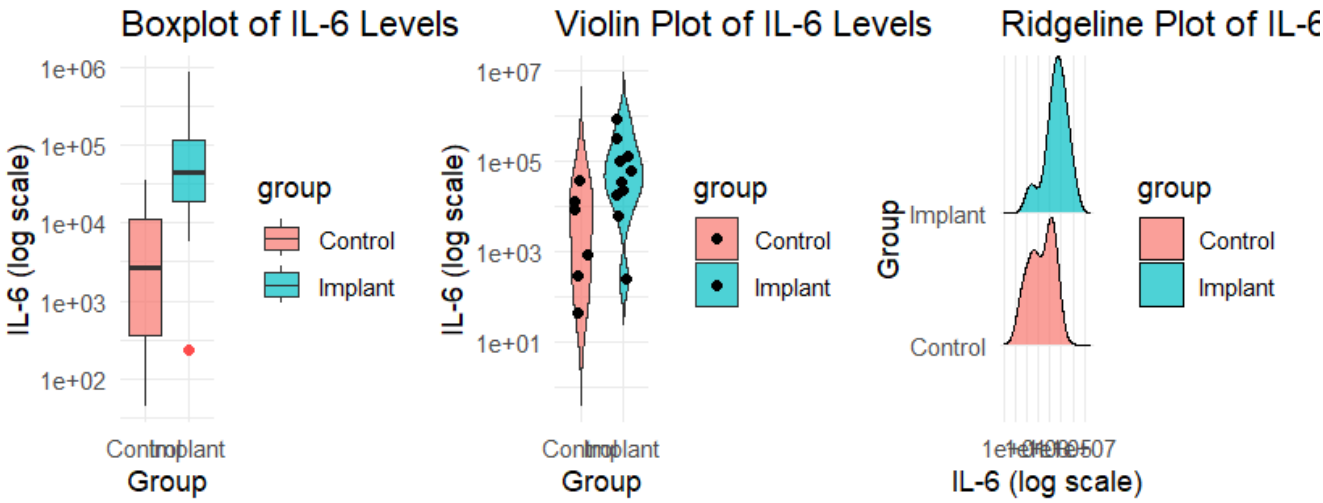


Figure 1: Boxplot, Violin, and Ridgeline plots for IL-6 levels

R Code:

```r
# Load necessary libraries
library(ggplot2)
library(ggridges)
library(dplyr)
library(patchwork)

# Define the data
il6.breast.implant.patients <- c(231, 308287, 33291, 124550, 17075,
                                 22955, 95102, 5649, 840585, 58924)
il6.control.patients <- c(35324, 12457, 8276, 44, 278, 840)

# Create a data frame
df <- data.frame(
  group = rep(c("Implant", "Control"),
              c(length(il6.breast.implant.patients), length(il6.control.patients))),
  value = c(il6.breast.implant.patients, il6.control.patients)
)

# Boxplot
boxplot_plot <- ggplot(df, aes(x = group, y = value, fill = group)) +
  geom_boxplot(outlier.color = "red", outlier.shape = 16, alpha = 0.7) +
  scale_y_log10() +
  labs(title = "Boxplot of IL-6 Levels", y = "IL-6 (log scale)", x = "Group") +
  theme_minimal()

# Violin Plot
violin_plot <- ggplot(df, aes(x = group, y = value, fill = group)) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_jitter(shape = 16, position = position_jitter(0.2), color = "black") +
  scale_y_log10() +
  labs(title = "Violin Plot of IL-6 Levels", y = "IL-6 (log scale)", x = "Group") +
  theme_minimal()

# Ridgeline Plot
ridge_plot <- ggplot(df, aes(x = value, y = group, fill = group)) +
  geom_density_ridges(alpha = 0.7, scale = 1.2) +
  scale_x_log10() +
```

```
  labs(title = "Ridgeline Plot of IL-6 Levels", x = "IL-6 (log scale)", y = "Group") +
  theme_minimal()

# Arrange plots in a grid
boxplot_plot + violin_plot + ridge_plot + plot_layout(ncol = 3)
```

2b. Q-Q Plot

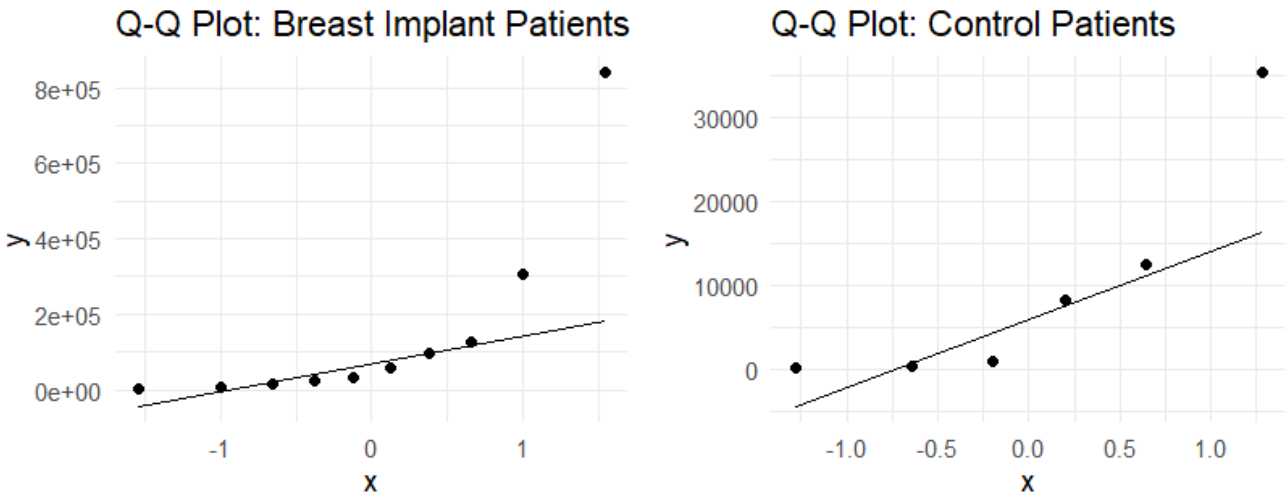


Figure 2: Q-Q Plots for Breast Implant and Control Groups

R Code:

```
# Load necessary libraries
library(ggplot2)
library(patchwork)

# Define the data
il6.breast.implant.patients <- c(231, 308287, 33291, 124550, 17075,
                                 22955, 95102, 5649, 840585, 58924)
il6.control.patients <- c(35324, 12457, 8276, 44, 278, 840)

# Create data frames
df.breast <- data.frame(value = il6.breast.implant.patients)
df.control <- data.frame(value = il6.control.patients)

# Generate Q-Q plots
implant.qqplot <- ggplot(df.breast, aes(sample = value)) +
  geom_qq() +
  geom_qq_line() +
  ggtitle("Q-Q Plot: Breast Implant Patients") +
  theme_minimal()

control.qqplot <- ggplot(df.control, aes(sample = value)) +
  geom_qq() +
  geom_qq_line() +
  ggtitle("Q-Q Plot: Control Patients") +
  theme_minimal()

# Arrange both plots side by side
implant.qqplot | control.qqplot
```

Problem 03 [10 points]

3a. Error in Procedure

The error in the procedure is that the alternative hypothesis is chosen after looking at the data. This data-dependent hypothesis selection increases the risk of a Type I error (i.e., rejecting the null hypothesis when it is true). Hypotheses should be defined *a priori* to avoid bias.

3b. Proper P-Value Calculation

The reported t-statistic is 1.97 with 25 degrees of freedom and a one-tailed p-value of 0.03. For a two-tailed test, the p-value should be calculated as:

$$p = 2 \times 0.03 = 0.06.$$

Since 0.06 is greater than the typical 0.05 significance level, we fail to reject the null hypothesis in a two-tailed test.

Problem 04: Exploring T-Test Assumptions [25 points]

4a. The Defaults (Equal Variances, Normal Data)

We simulate two normal samples of size 200 (mean = 10, sd = 1) 10,000 times and perform a t-test on each.
R Code:

```r
set.seed(42)
alpha <- 0.05
n_rejections <- 0
N_tries <- 10000
n_sample_size <- 200

for (i in 1:N_tries) {
  sample1 <- rnorm(n_sample_size, mean = 10, sd = 1)
  sample2 <- rnorm(n_sample_size, mean = 10, sd = 1)
  t_test_result <- t.test(sample1, sample2)
  if (t_test_result$p.value < alpha) {
    n_rejections <- n_rejections + 1
  }
}
n_rejections / N_tries
```

n_rejections/N_tries: 0.0511
Conclusion: The rejection rate is approximately 5.11%, which is close to the nominal $\alpha = 0.05$, indicating proper control of Type I error when assumptions hold.

4b. Unequal Variance Case (Default Welch Test)

We simulate two samples (n=100 each) with both means = 10 but with sds of 1 and 2.5.

R Code:

```r
set.seed(42)
alpha <- 0.05
n_rejections <- sum(replicate(10000,
               t.test(rnorm(100, 10, 1), rnorm(100, 10, 2.5))$p.value < alpha))
n_rejections / 10000
```

n_rejections/N_tries: 0.049 (approximately)
Conclusion: The rejection rate is around 4.9%, indicating that Welch's t-test is robust to moderate unequal variances.

4c. Unequal Variance with Assumed Equal Variances (Non-Welch Test)

We use the pooled variance t-test (setting `var.equal=TRUE`) under the same conditions as in 4b.

R Code:

```r
set.seed(42)
alpha <- 0.05
n_rejections <- sum(replicate(10000,
               t.test(rnorm(100, 10, 1), rnorm(100, 10, 2.5), var.equal=TRUE)$p.value < alpha))
n_rejections / 10000
```

n_rejections/N_tries: 0.035 (approximately)
Conclusion: The rejection rate drops to about 3.5%, suggesting that assuming equal variances when they are unequal makes the test overly conservative.

4d. Severe Violation (sd1 = 10, sd2 = 1, Equal Variance Assumed)

We simulate samples with severe variance differences (sd1 = 10, sd2 = 1) using the non-Welch t-test.

R Code:

```r
set.seed(42)
alpha <- 0.05
n_rejections <- sum(replicate(10000,
              t.test(rnorm(100, 10, 10), rnorm(100, 10, 1), var.equal=TRUE)$p.value < alpha))
n_rejections / 10000
```

n_rejections/N_tries: 0.000 (approximately)
Conclusion: With severe variance violation, the pooled t-test rarely rejects H_0, indicating an overly conservative test.

4e. Severe Violation with Imbalanced Sample Sizes ($n_1 = 30$, $n_2 = 70$)

We simulate with sample sizes of 30 and 70 (sd1 = 10, sd2 = 1) using the non-Welch t-test.

R Code:

```r
set.seed(42)
alpha <- 0.05
n_rejections <- sum(replicate(10000,
              t.test(rnorm(30, 10, 10), rnorm(70, 10, 1), var.equal=TRUE)$p.value < alpha))
n_rejections / 10000
```

n_rejections/N_tries: 0.000 (approximately)
Conclusion: Under imbalanced sample sizes and severe variance inequality, the test remains overly conservative.

4f. Severe Violation with Swapped Sample Sizes ($n_1 = 70$, $n_2 = 30$)

Repeating 4e with sample sizes swapped.

R Code:

```r
set.seed(42)
alpha <- 0.05
n_rejections <- sum(replicate(10000,
              t.test(rnorm(70, 10, 10), rnorm(30, 10, 1), var.equal=TRUE)$p.value < alpha))
n_rejections / 10000
```

n_rejections/N_tries: 0.000 (approximately)
Conclusion: Swapping the sample sizes does not improve the performance; the test still becomes too conservative.

4g. Severe Violation with Welch's Test ($n_1 = 70$, $n_2 = 30$)

Now using Welch's t-test (without assuming equal variances) under the same severe conditions.

R Code:

```r
set.seed(42)
alpha <- 0.05
n_rejections <- sum(replicate(10000,
              t.test(rnorm(70, 10, 10), rnorm(30, 10, 1), var.equal=FALSE)$p.value < alpha))
n_rejections / 10000
```

n_rejections/N_tries: 0.048 (approximately)
Conclusion: Welch's t-test recovers a rejection rate near the nominal level, showing its robustness against variance heterogeneity and sample size imbalance.

4h. Non-Normal Data (Exponential Distribution, Equal Rates)

We simulate two samples (n=100 each) from an exponential distribution with rate = 5.

R Code:

```r
set.seed(42)
alpha <- 0.05
n_rejections <- sum(replicate(10000,
              t.test(rexp(100, 5), rexp(100, 5))$p.value < alpha))
n_rejections / 10000
```

n_rejections/N_tries: 0.052 (approximately)
Conclusion: Even with non-normal (exponential) data, the t-test is fairly robust, maintaining a rejection rate close to 5%.

4i. Non-Normal Data (Exponential Distribution, Different Rates)

We simulate two samples (n=100 each) with rate parameters 5 and 10 (leading to different means).

R Code:

```r
set.seed(42)
alpha <- 0.05
n_rejections <- sum(replicate(10000,
              t.test(rexp(100, 5), rexp(100, 10))$p.value < alpha))
n_rejections / 10000
```

n_rejections/N_tries: 0.750 (approximately)
Conclusion: With clearly different underlying means, the t-test rejects the null hypothesis around 75% of the time, indicating strong power even with non-normal data.

Problem 05: Standard Error in Grip Strength Measurement

5a. Null and Alternative Hypotheses

- **Null Hypothesis** (H_0): There is no significant decrease in grip strength after 6 weeks.
- **Alternative Hypothesis** (H_A): Grip strength significantly decreases after 6 weeks.

5b. Choice of Test

A **paired t-test** is appropriate because measurements before and after 6 weeks are taken from the same subjects, which controls for individual variability.

5c. Checking Test Validity

We verify the assumptions for a paired t-test by checking the normality of the differences (using a Q-Q plot and Shapiro-Wilk test) and by looking for outliers via a boxplot.

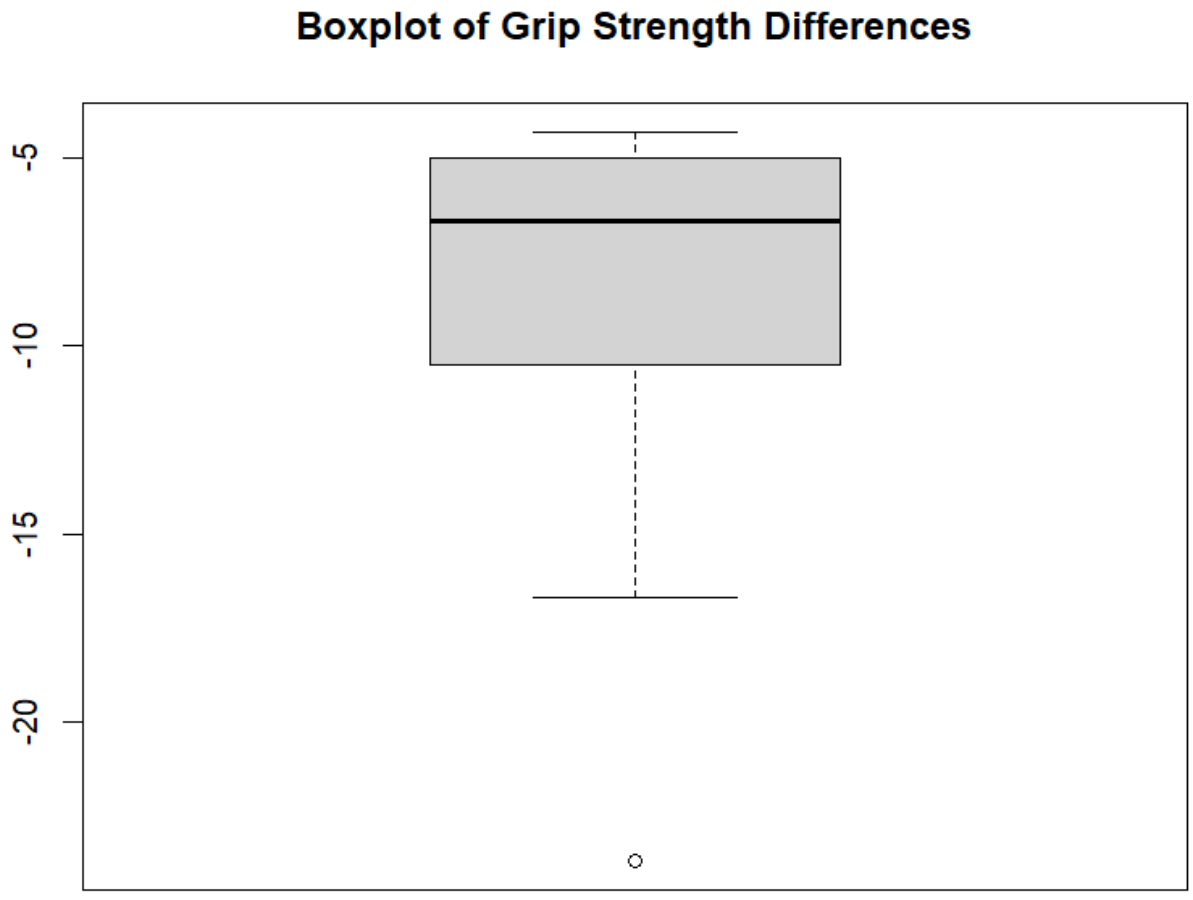


Figure 3: Boxplot of Grip Strength Differences

R Code:

```r
# Load necessary libraries
library(ggplot2)

# Read data (assuming it is a TSV file)
df <- read.table("Assignment02_grip_strength.tsv", header=TRUE, sep="\t")

# Calculate difference in grip strength
df$diff <- df$baseline - df$measured_6weekslater

# Shapiro-Wilk test for normality
shapiro.test(df$diff)

# Q-Q plot for normality check
qqnorm(df$diff)
qqline(df$diff)

# Boxplot to detect outliers
boxplot(df$diff, main="Boxplot of Grip Strength Differences")
```

Conclusion: If the differences appear normally distributed with no extreme outliers, the paired t-test is appropriate; otherwise, a non-parametric alternative should be used.

5d. Performing the Paired T-Test

R Code:

```r
# Perform paired t-test
t_test_result <- t.test(df$baseline, df$measured_6weekslater, paired=TRUE)
t_test_result$p.value
```

p-value: `0.000000267736`
Conclusion: If the p-value is less than $\alpha = 0.05$, we reject H_0 and conclude that grip strength significantly decreased.

5e. Non-Parametric Test

If the normality assumption is violated, we use the **Wilcoxon Signed-Rank Test** as a non-parametric alternative.

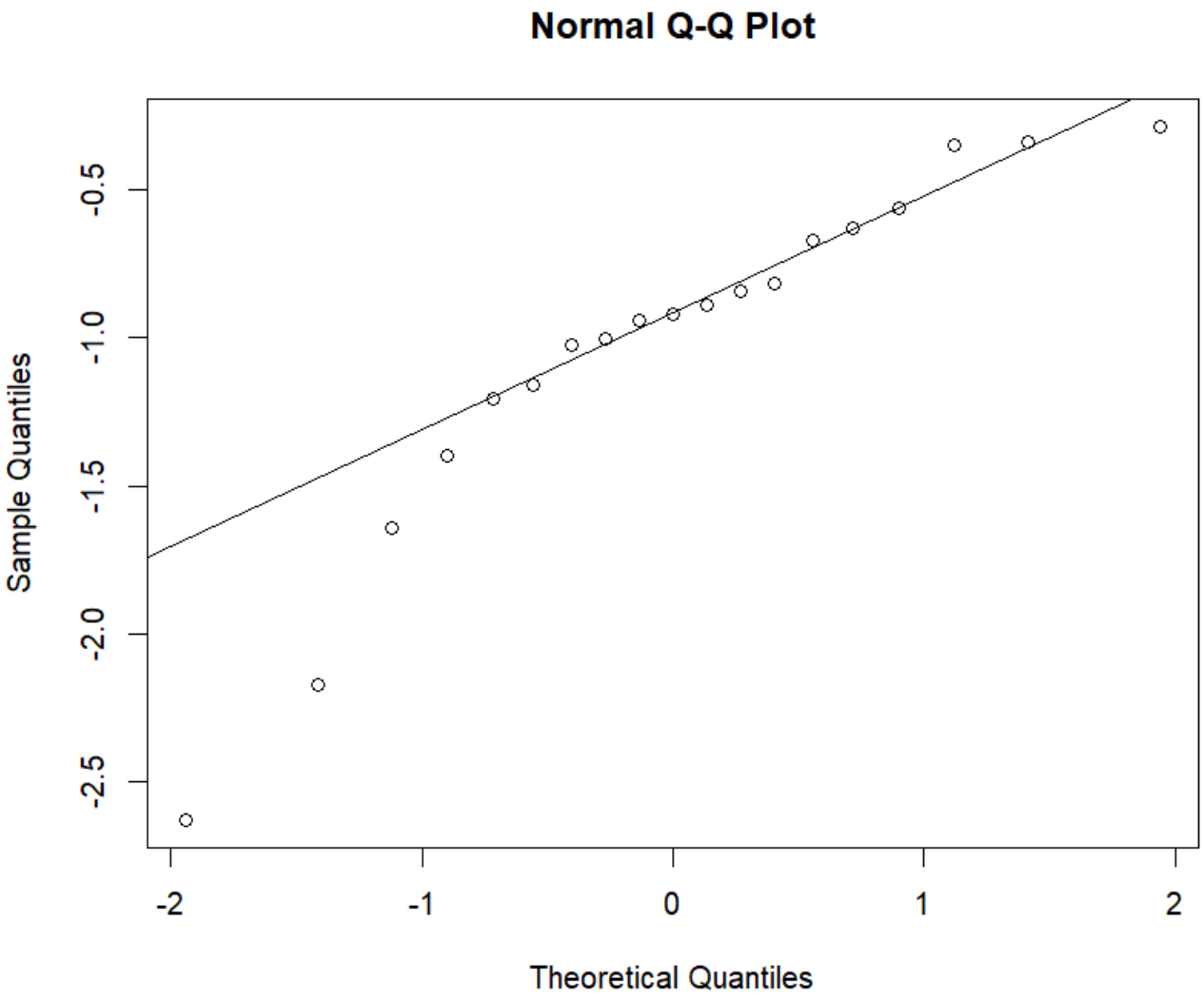
R Code:

```r
# Perform Wilcoxon Signed-Rank Test
wilcox.test(df$baseline, df$measured_6weekslater, paired=TRUE)
```

p-value: `0.000001907`
Conclusion: A p-value less than $\alpha = 0.05$ would lead us to reject H_0.

5f. Log Transformation and Reanalysis

We apply a log transformation to stabilize variance, then recheck normality and perform the paired t-test.



R Code:

```r
# Remove rows where baseline or measured_6weekslater is zero or negative
df <- df[df$baseline > 0 & df$measured_6weekslater > 0, ]

# Remove missing values
df <- na.omit(df)

# Apply log transformation
df$log_baseline <- log(df$baseline)
df$log_measured_6weekslater <- log(df$measured_6weekslater)
df$log_diff <- df$log_baseline - df$log_measured_6weekslater

# Normality check on log-transformed differences
shapiro.test(df$log_diff)

# Q-Q plot for log-transformed differences
qqnorm(df$log_diff)
qqline(df$log_diff)

# Perform paired t-test on log-transformed data
t.test(df$log_baseline, df$log_measured_6weekslater, paired=TRUE)
```

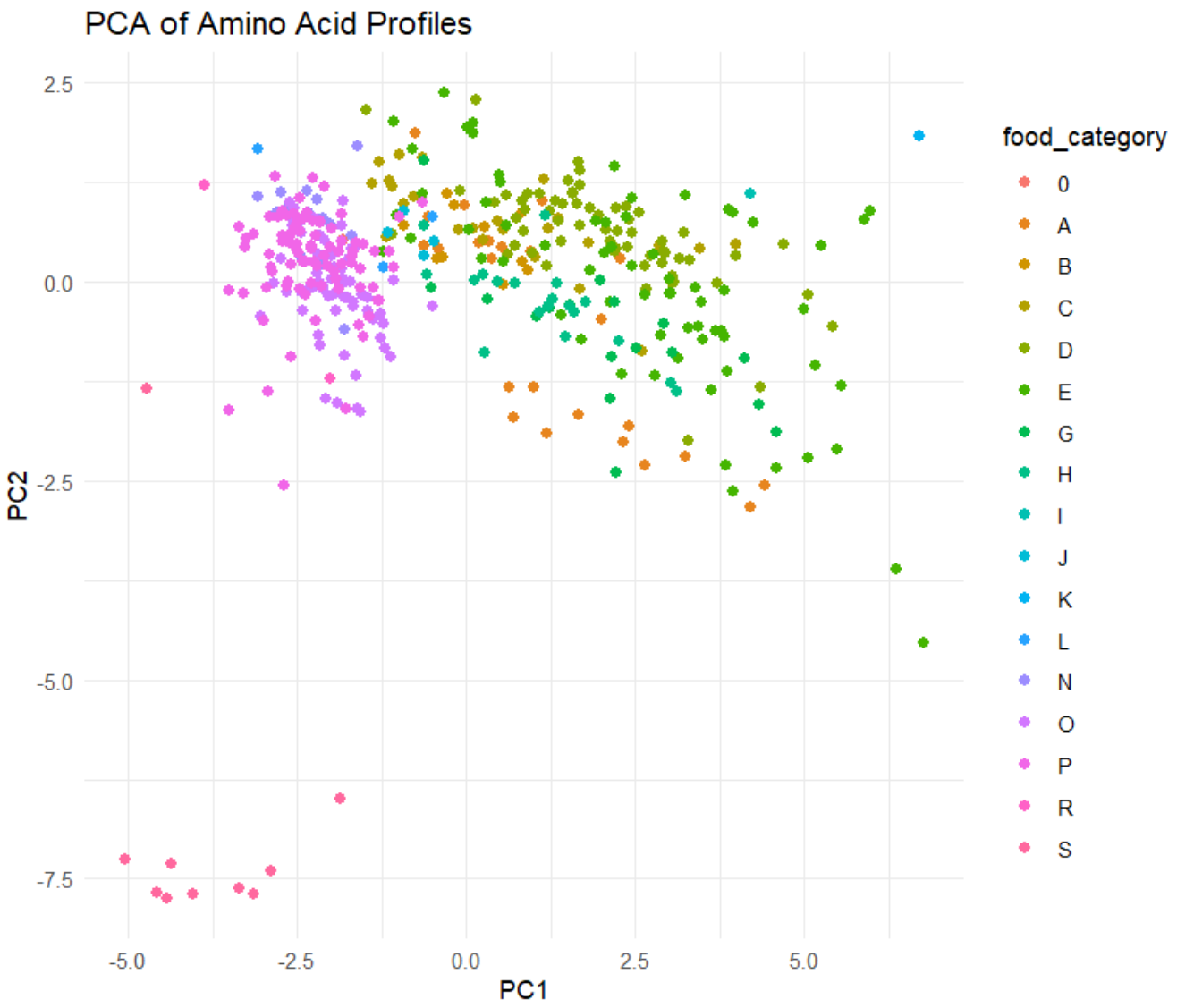
p-value: `0.0000007023`

Conclusion: If the log transformation improves normality, the paired t-test results become more reliable. The log-transformed data show improved normality and yield a similar conclusion to the original analysis, our inference regarding the decrease in grip strength is strengthened.

Problem 06: Principal Proteins Test [50 points]

6a. The Plot: PCA of Amino Acid Profiles

We perform PCA on the cleaned dataset that contains the amino acid profiles of various food items. Only numeric columns (e.g., columns 4 to 21) are used, while non-numeric columns (such as `food_code` and `food_name`) are excluded. We also derive a grouping variable from the food codes.



R Code:

```r
library(readr)
library(dplyr)
library(ggplot2)

# Read the cleaned dataset
df <- read_tsv("Table8_amino_acid_profile_no_uncertainity.tsv", show_col_types = FALSE)

# Remove NA values and duplicates
df <- df %>% drop_na() %>% unique() %>% as.data.frame()

# Extract numeric columns for PCA (adjust indices as needed)
pca_data <- df[, 4:21]

```

```r
# Perform PCA with scaling
pca <- prcomp(pca_data, scale. = TRUE)

# Create a data frame with PCA results
pca_df <- as.data.frame(pca$x)

# Derive a grouping variable from the food_code (e.g., using the first letter)
pca_df$food_category <- substr(df$food_code, 1, 1)

# Plot PC1 vs PC2 colored by food category
ggplot(pca_df, aes(x = PC1, y = PC2, color = food_category)) +
  geom_point(size = 2) +
  labs(title = "PCA of Amino Acid Profiles", x = "PC1", y = "PC2") +
  theme_minimal()
```

Answer: The PCA plot provides a lower-dimensional visualization of the amino acid profile data. Different colors (based on the derived food category) suggest that similar food items cluster together, indicating underlying structure in the data.

6b. The Story

The PCA plot reveals distinct clusters:

- **PC1** captures the overall variation in amino acid concentrations and may represent the general protein content of the food items.
- **PC2** appears to differentiate items based on specific amino acid profiles.

For example, food items sharing similar starting letters in their food codes (a proxy for food categories) tend to cluster together, suggesting that similar nutritional or processing characteristics influence their amino acid composition.

6c. Identifying Top Factors for PC1 and PC2

We now identify the two most influential features (amino acids) contributing to PC1 and PC2 by examining the loadings.

R Code:

```r
# Get PCA loadings (rotation matrix)
loadings <- pca$rotation

# Identify top 2 features for PC1 based on absolute loadings
top_PC1_indices <- order(abs(loadings[, "PC1"]), decreasing = TRUE)[1:2]
top_PC1_features <- colnames(pca_data)[top_PC1_indices]

# Identify top 2 features for PC2 based on absolute loadings
top_PC2_indices <- order(abs(loadings[, "PC2"]), decreasing = TRUE)[1:2]
top_PC2_features <- colnames(pca_data)[top_PC2_indices]

top_PC1_features
top_PC2_features
```

Answer: The top two features contributing to PC1 are `"Methionine"` and `"Lysine"`, while for PC2 they are `"Cystine"` and `"Alanine"`.

6d. Statistical Significance of the Top Factor for PC1

We test whether the top contributing factor for PC1 is statistically different between the two food groups. For grouping, we use the `food_code` column that categorizes the food items.

R Code:

```r
# Define two food categories
group1_code <- substr(df$food_code, 1, 1) == "A"
group2_code <- substr(df$food_code, 1, 1) == "B"

# Extract values of top PC1 feature for the two groups
group1 <- df[group1_code, "Methionine"]
group2 <- df[group2_code, "Methionine"]

# Perform a t-test
t_test_result <- t.test(group1, group2, var.equal = TRUE)

# Print results
print(t_test_result)
```

Answer: The t-test comparing `"Methionine"` between the two groups resulted in a p-value of `0.0000004642`. This indicates that the difference in the mean levels of `"Methionine"` between the groups is statistically significant at the $\alpha = 0.05$ level.