

Exam1.STA106

Darren

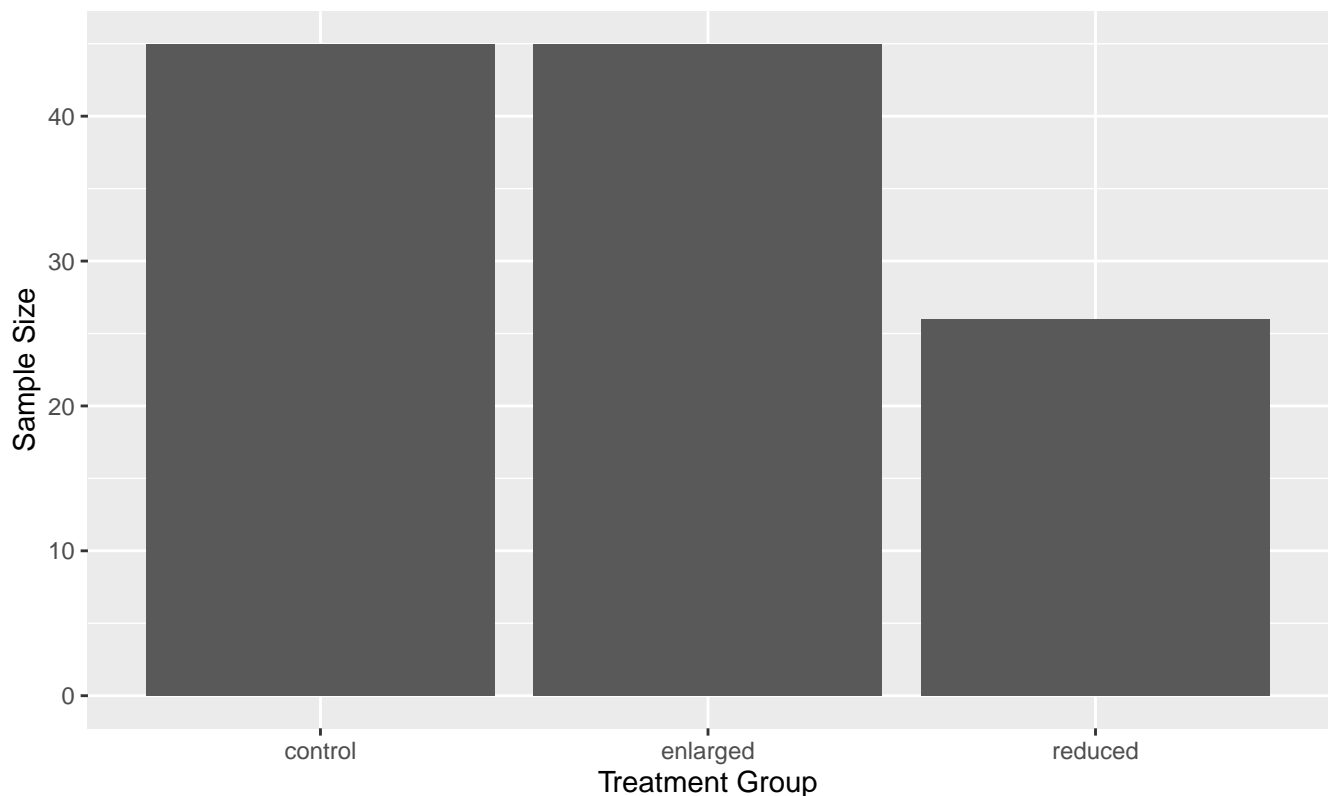
2025-04-27

I. Introduction. State the question you are trying to answer, why it is a question of interest (why might we be interested in the answer), and what approach you are going to take (just the name of the approach).

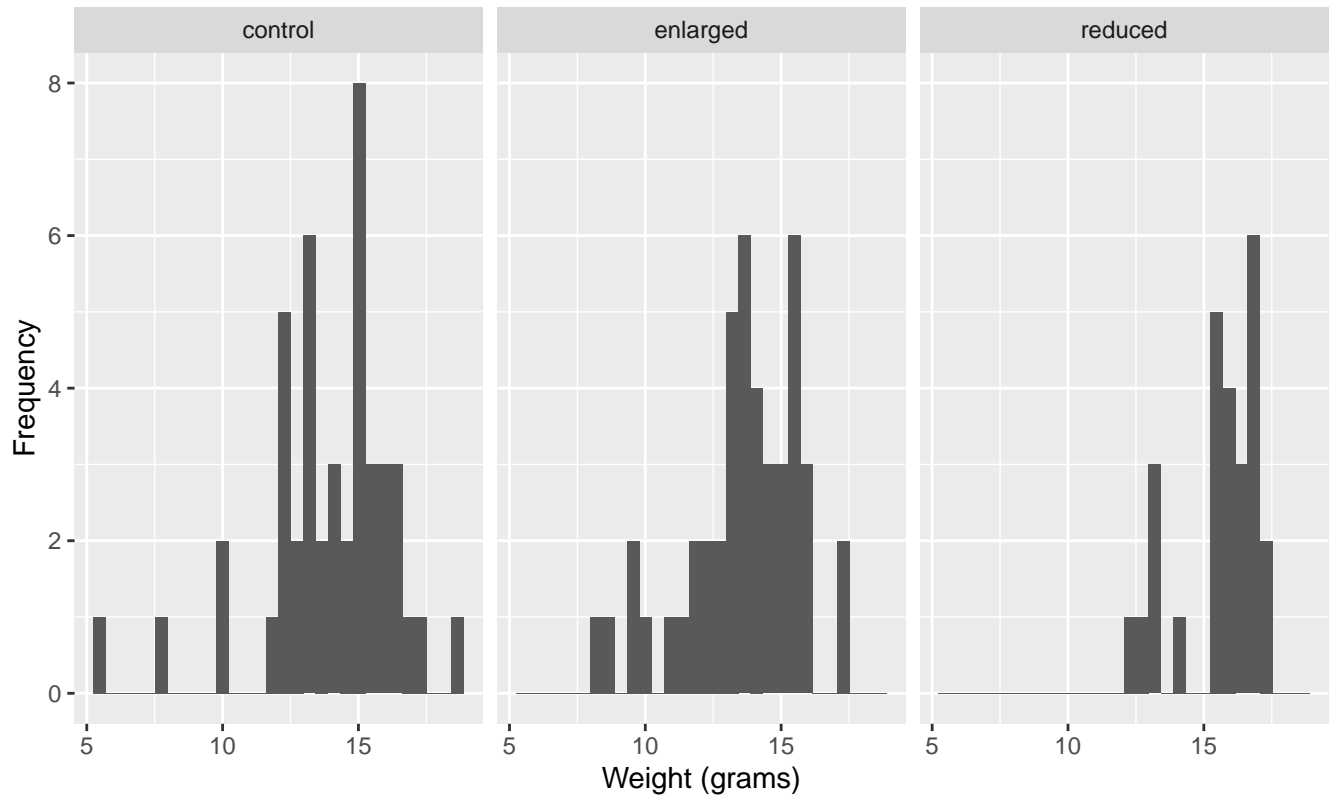
II. Summary of your data. This should include things like plots (histograms, boxplots) including the interpretation of the plots, and summary values such as sample means and standard deviations, etc. You should have an idea about the trend of the data from this section.

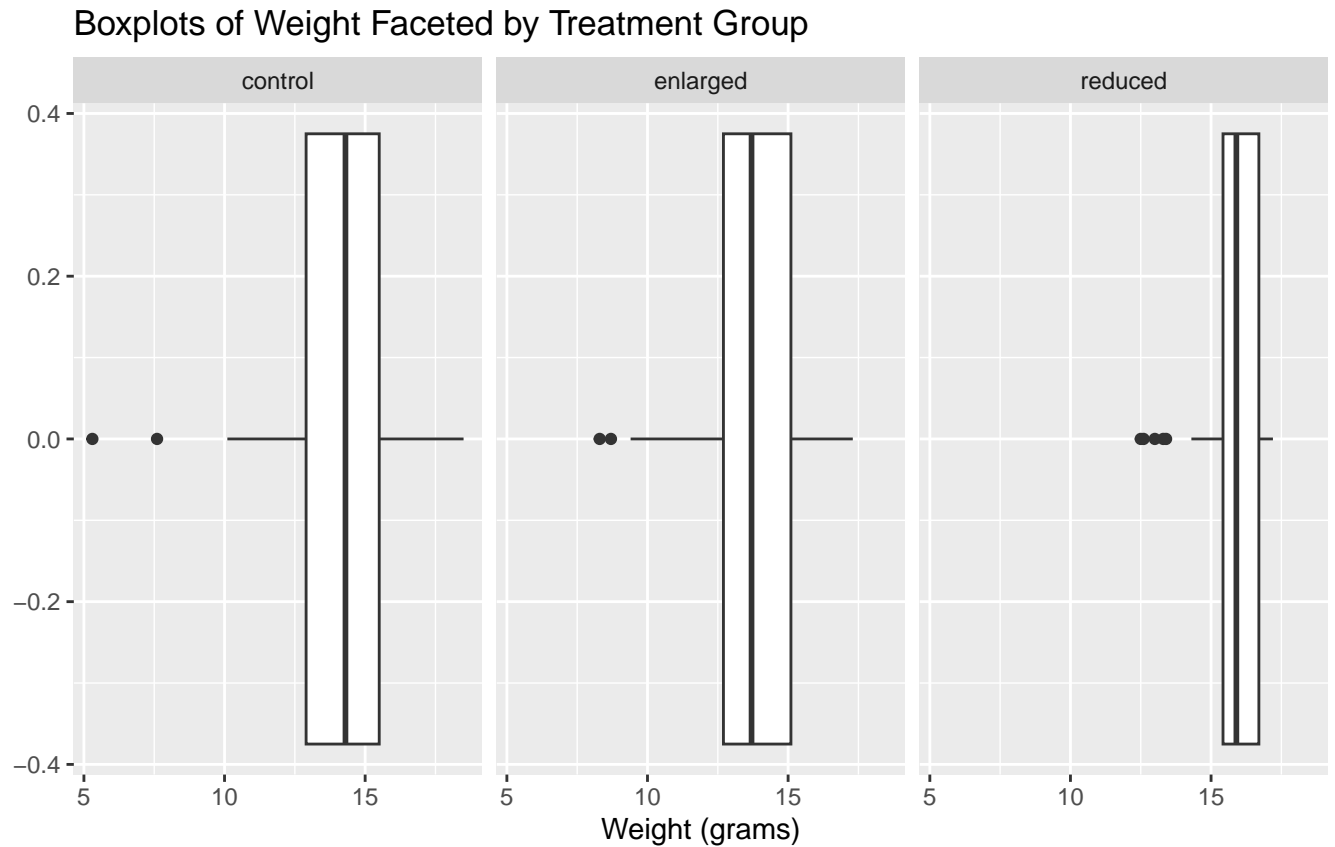
##	Treatment	Mean	SD	n
## 1	control	13.9244444444444	2.41963141028053	45
## 2	enlarged	13.5155555555556	2.10399571762824	45
## 3	reduced	15.5692307692308	1.45925165225721	26
## 4	overall	14.1344827586207	2.24259567963306	116

Bar Graph of Treatment Groups



Histograms of Weight Faceted by Treatment Group

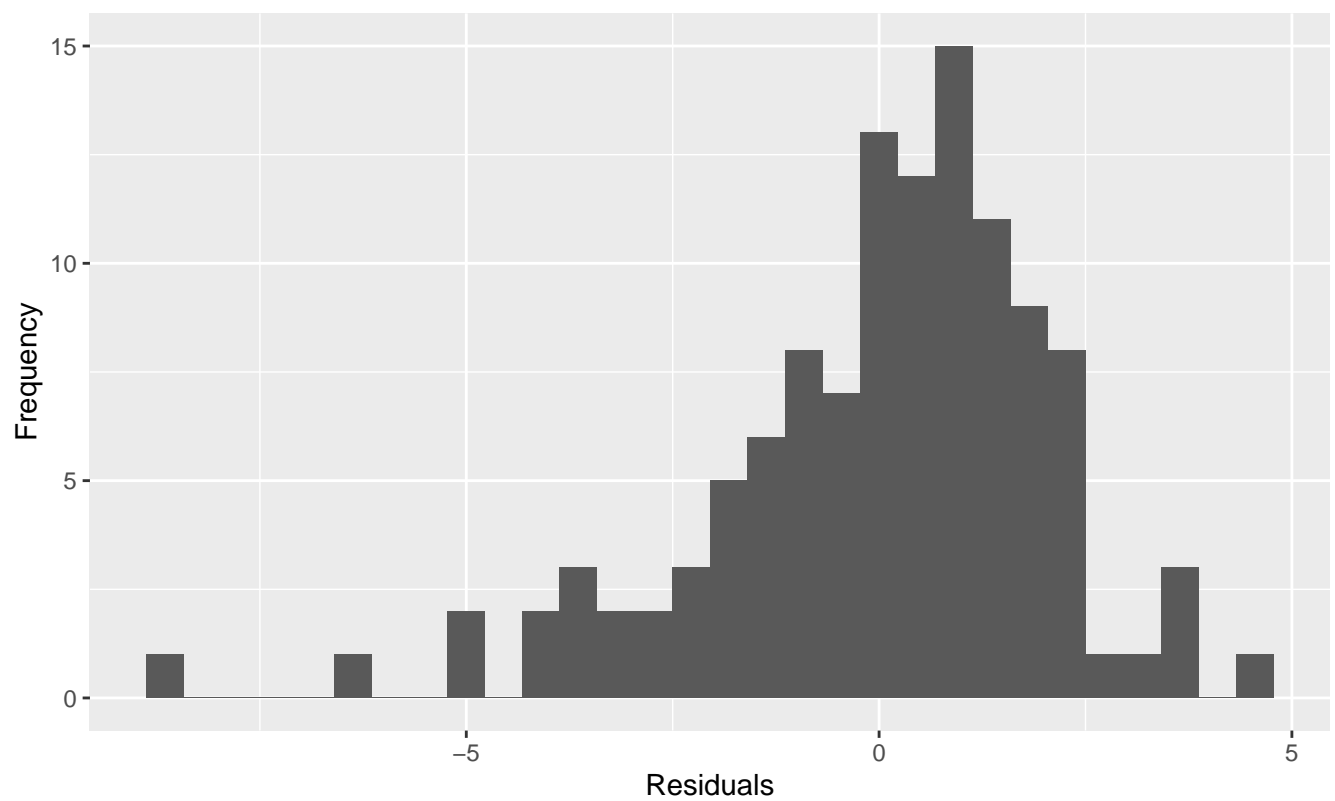




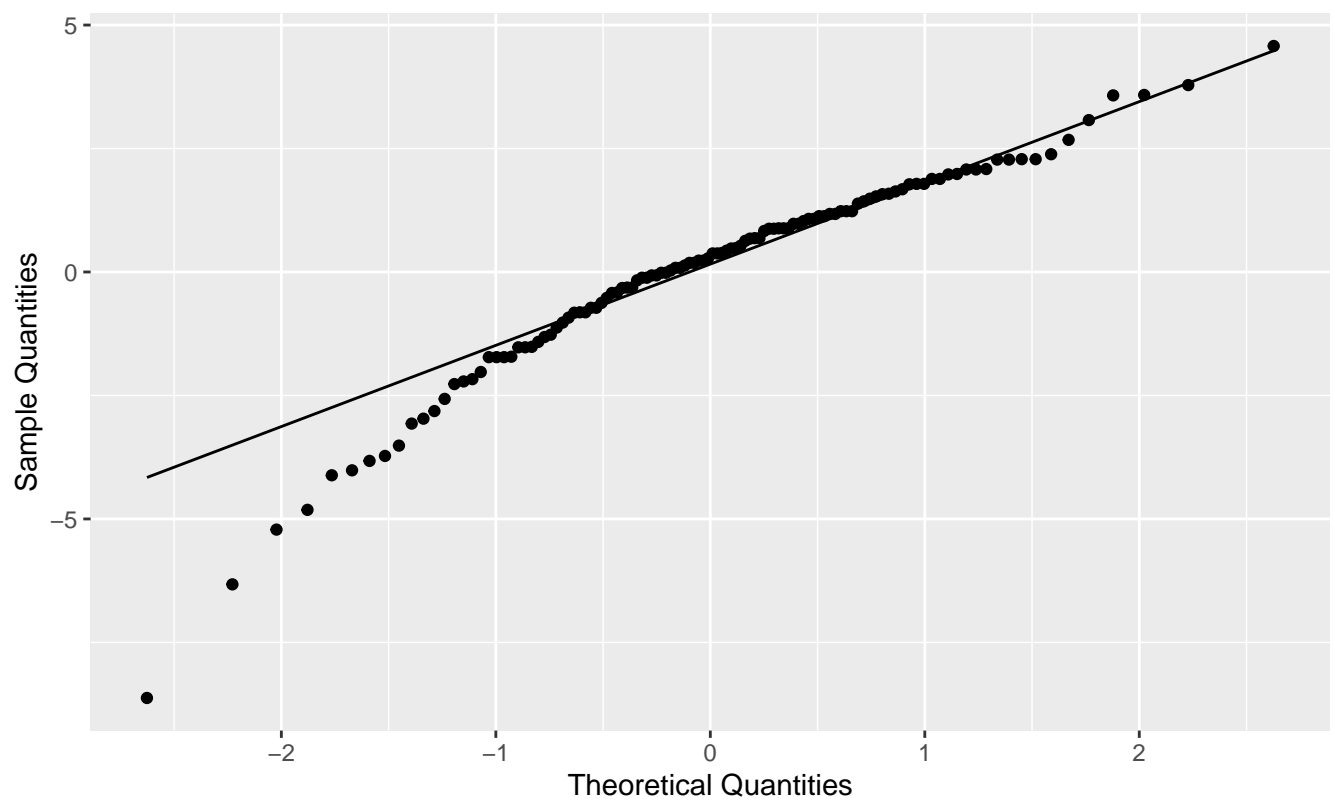
III. Diagnostics. You should discuss your assumptions here, and if you believe they are violated. Perform diagnostics for the model. Remove outliers if necessary. You do not need to do transformation of variables.

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Treatment    2   72.7   36.37   8.129 0.000503 ***
## Residuals  113  505.6    4.47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Histogram of Residuals (Pre-Outlier-Removal)



Normal QQ Plot of Residuals (Pre-Outlier-Removal)



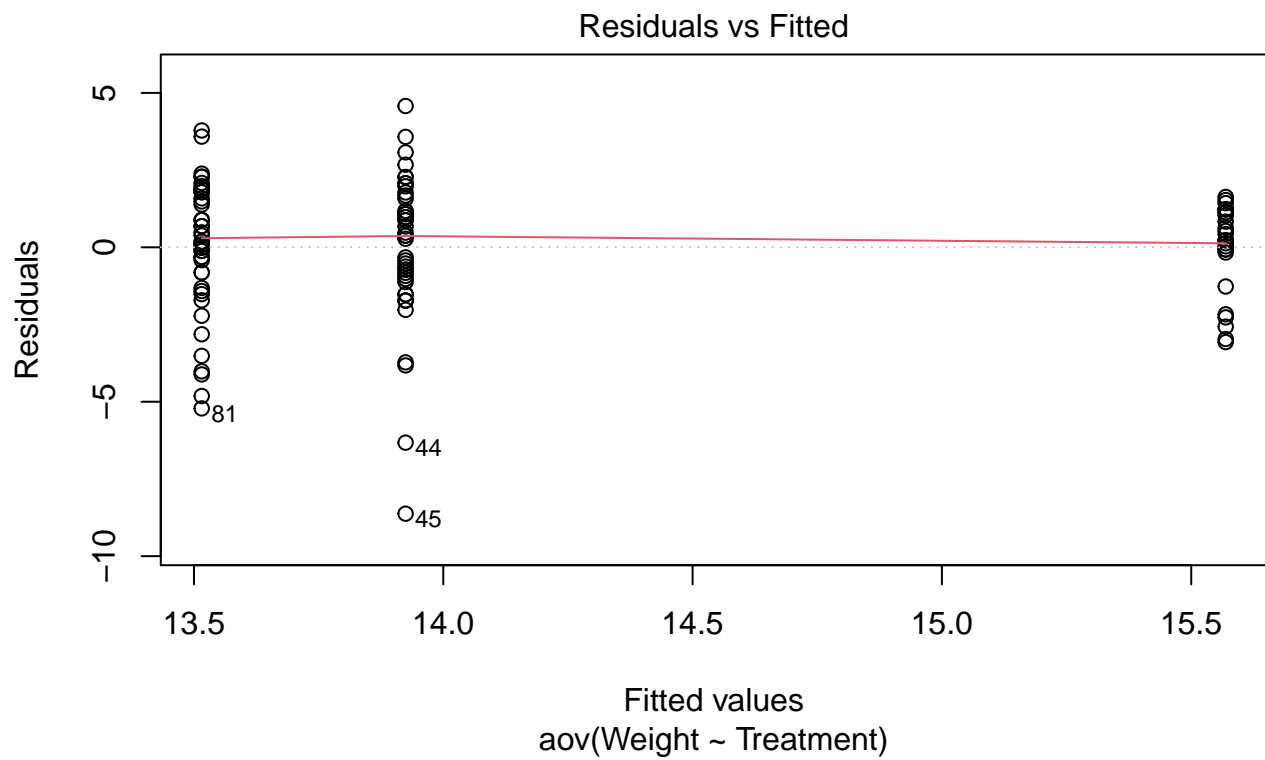
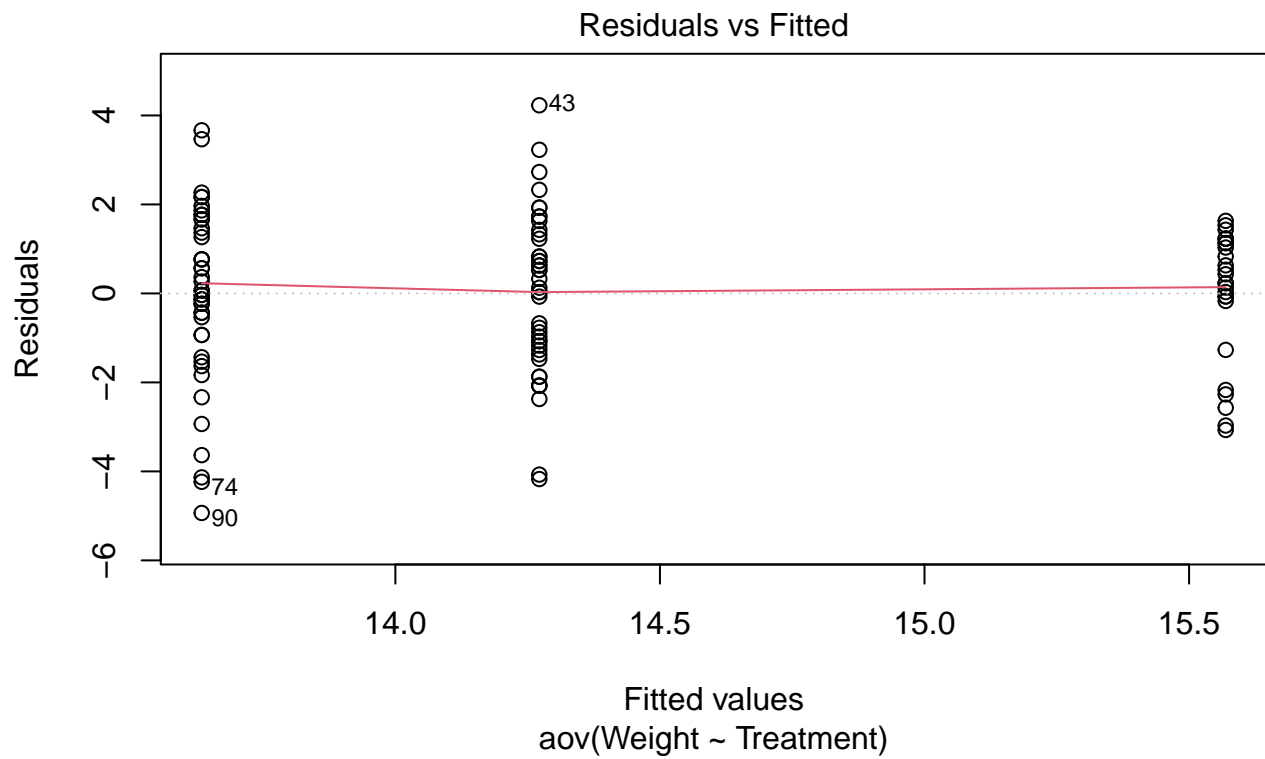
```
##
## Shapiro-Wilk normality test
##
## data:  initial_model$residuals
## W = 0.94055, p-value = 0.00006245

##
## Brown-Forsythe Test (alpha = 0.05)
## -----
## data : Weight and Treatment
##
## statistic : 9.155928
## num df : 2
## denom df : 111.159
## p.value : 0.0002085043
##
## Result : Difference is statistically significant.
## -----

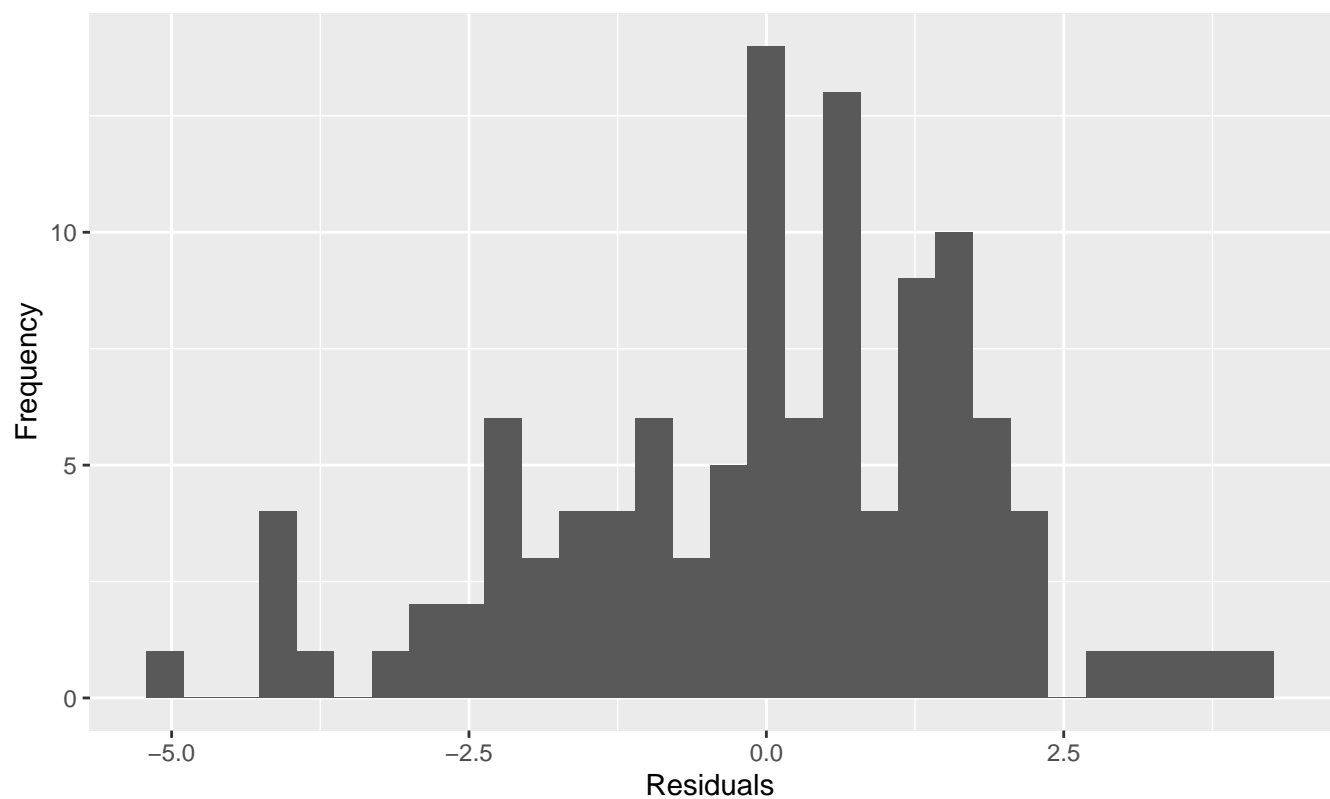
## 44 45 81
## 44 45 81

## Treatment Weight
## 44 control 7.6
## 45 control 5.3
## 81 enlarged 8.3

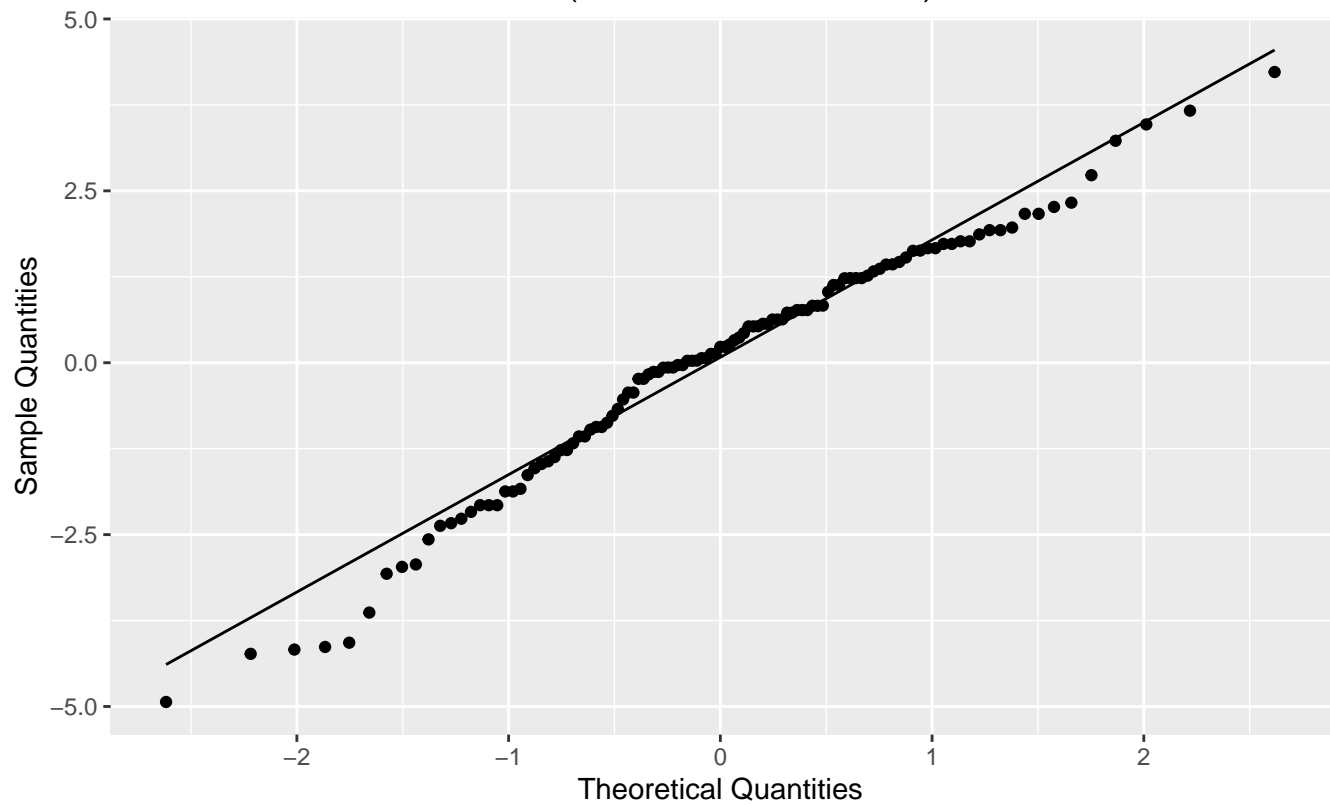
## Df Sum Sq Mean Sq F value Pr(>F)
## Treatment 2 61.4 30.687 9.423 0.000167 ***
## Residuals 110 358.2 3.257
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Histogram of Residuals (Post-Outlier-Removal)



Normal QQ Plot of Residuals (Post-Outlier-Removal)



```
##
## Shapiro-Wilk normality test
##
## data:  new_model$residuals
## W = 0.97568, p-value = 0.0371

##
## Brown-Forsythe Test (alpha = 0.05)
## -----
## data : Weight and Treatment
##
## statistic : 10.15092
## num df : 2
## denom df : 108.5019
## p.value : 0.00009094607
##
## Result : Difference is statistically significant.
## -----
```

IV. Analysis & Interpretation. Report back the model fit, confidence intervals, test-statistic/s, and p-value/s, nulls and alternatives, power calculations, etc. Be sure that you organize your work and write your results in full sentences where possible. State your conclusion, and what inference you may draw from your corresponding tests or confidence intervals. These should all be in terms of your problem.

```
## Treatment      Mean      SD      n
## 1 control 14.2720930232558 1.8128280434035 43
## 2 enlarged 13.6340909090909 1.97047310832021 44
## 3 reduced 15.5692307692308 1.45925165225721 26
## 4 overall 14.3221238938053 1.93555706789666 113

## [1] 0.9766715 2.5064111

## Estimate Lower Upper
## reduced 15.5692308 14.867865 16.2705966
## control - enlarged 0.6380021 -0.128883 1.4048872
## control - reduced -1.2971377 -2.185592 -0.4086839
```

V. Conclusion. Summarize briefly your findings. Here you do not have to re-iterate your numeric values, but summarize all relevant conclusions from your initial introduction.

R Appendix

```
# Setup
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.align =
"center", fig.width = 7)
options(scipen = 999) #Remove the scientific notation

# Load libraries
library(dplyr)
library(ggplot2)
```



```

library(onewaytests)
library(MASS)

# Import the dataset
sparrow <- read.csv("Sparrow.csv")
# II. Exploratory Data Analysis
# Calculate summary statistics for each treatment group
data.frame(sparrow %>%
  group_by(Treatment) %>%
  summarize(Mean = mean(Weight, na.rm = TRUE),
            SD = sd(Weight, na.rm = TRUE),
            n = n()) %>%
  rbind(c(Treatment = 'overall',
          Mean = mean(sparrow$Weight),
          SD = sd(sparrow$Weight),
          n = length(sparrow$Weight)))

# Create bar graph for each treatment group
sparrow %>%
  ggplot(mapping = aes(x = Treatment)) +
  geom_bar() +
  labs(title = "Bar Graph of Treatment Groups",
       x = "Treatment Group",
       y = "Sample Size")

# Create histograms for each treatment group
sparrow %>%
  ggplot(mapping = aes(x = Weight)) +
  geom_histogram() +
  facet_wrap(~Treatment) +
  labs(title = "Histograms of Weight Faceted by Treatment Group",
       x = "Weight (grams)",
       y = "Frequency")

# Create boxplots for each treatment group
sparrow %>%
  ggplot(mapping = aes(x = Weight)) +
  geom_boxplot() +
  facet_wrap(~Treatment) +
  labs(title = "Boxplots of Weight Faceted by Treatment Group",
       x = "Weight (grams)")
# III. Initial Model Diagnostics
# Pre-outlier-removal analysis
# Create the model
initial_model <- aov(Weight ~ Treatment, data = sparrow)
summary(initial_model)

# Create histogram to show residual distribution
initial_model %>%
  ggplot(mapping = aes(x = .resid)) +
  geom_histogram() +
  labs(title = "Histogram of Residuals (Pre-Outlier-Removal)",
       x = "Residuals",

```

```

    y = "Frequency")

# Create QQ plot for residual normality
initial_model %>%
  ggplot(mapping = aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Normal QQ Plot of Residuals (Pre-Outlier-Removal)",
        x = "Theoretical Quantities",
        y = "Sample Quantities")

# Perform Shapiro-Wilks Test for Normality
shapiro.test(initial_model$residuals)
# Perform Brown-Forsythe Test for Constant Variance
bf.test(Weight ~ Treatment, data = sparrow)

# Check for any outliers
stud_resids <- studres(initial_model)
outliers <- which(abs(stud_resids) > 2.5)
outliers

# View and Remove outliers
sparrow[outliers, ]
sparrow_clean <- sparrow[-outliers, ]
# Post Outlier Removal
# Refit model on cleaned data
new_model <- aov(Weight ~ Treatment, data = sparrow_clean)
summary(new_model)
plot(new_model, which = 1)
plot(initial_model, which = 1)
# Create histogram to show residual distribution
new_model %>%
  ggplot(mapping = aes(x = .resid)) +
  geom_histogram() +
  labs(title = "Histogram of Residuals (Post-Outlier-Removal)",
        x = "Residuals",
        y = "Frequency")

# Create QQ plot for residual normality
new_model %>%
  ggplot(mapping = aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Normal QQ Plot of Residuals (Post-Outlier-Removal)",
        x = "Theoretical Quantities",
        y = "Sample Quantities")

# Perform Shapiro-Wilks Test for Normality
shapiro.test(new_model$residuals)
# Perform Brown-Forsythe Test for Constant Variance
bf.test(Weight ~ Treatment, data = sparrow_clean)
# Calculate new summary statistics for each treatment group
data.frame(sparrow_clean %>%

```

```

group_by(Treatment) %>%
  summarize(Mean = mean(Weight, na.rm = TRUE),
            SD = sd(Weight, na.rm = TRUE),
            n = n()) %>%
  rbind(c(Treatment = 'overall',
          Mean = mean(sparrow_clean$Weight),
          SD = sd(sparrow_clean$Weight),
          n = length(sparrow_clean$Weight)))

# Define function to calculate power
calculate_power = function(ybar, ni, MSE, alpha){
  a = length(ybar) # Finds a
  nt = sum(ni) # Finds the overall sample size
  overall.mean = sum(ni*ybar)/nt # Finds the overall mean
  phi = (1/sqrt(MSE))*sqrt(sum(ni*(ybar - overall.mean)^2)/a) # Finds the books value of phi
  phi.star = a * phi^2 # Finds the value of phi we will use for R
  Fc = qf(1-alpha, a-1, nt-a) # The critical value of F, use in R's function
  power = 1 - pf(Fc, a-1, nt-a, phi.star) # The power, calculated using a non-central F
  return(c(power, phi))
}

# Calculate the power for the test with cleaned data
group_means <- by(sparrow_clean$Weight, sparrow_clean$Treatment, mean)
group_sizes <- by(sparrow_clean$Weight, sparrow_clean$Treatment, length)
MSE <- anova(new_model)[2, 3]
calculate_power(group_means, group_sizes, MSE, 0.05)

# Create 95% CI for each desired metric
multiplier = qt(1 - 0.05 / 2, 113 - 3)
data.frame(
  Estimate = group_means[3],
  Lower = group_means[3] - multiplier * sqrt(MSE/group_sizes[3]),
  Upper = group_means[3] + multiplier * sqrt(MSE/group_sizes[3])
) %>%
  rbind("control - enlarged" =
    c(Estimate = group_means[1] - group_means[2],
      Lower = (group_means[1] - group_means[2]) - multiplier
        * sqrt(MSE*(1/group_sizes[1] + 1/group_sizes[2])),
      Upper = (group_means[1] - group_means[2]) + multiplier
        * sqrt(MSE*(1/group_sizes[1] + 1/group_sizes[2])))
  ) %>%
  rbind("control - reduced" =
    c(Estimate = group_means[1] - group_means[3],
      Lower = (group_means[1] - group_means[3]) - multiplier
        * sqrt(MSE*(1/group_sizes[1] + 1/group_sizes[3])),
      Upper = (group_means[1] - group_means[3]) + multiplier
        * sqrt(MSE*(1/group_sizes[1] + 1/group_sizes[3])))
  )
)
tinytex::install_tinytex()

```