Post-weaning diarrhea I statistics report

Course: BMD407 Biomedical Statistics-Spring 2023

Submitted to: Dr. Mohamed Mysara and Eng. Mariam Oweda

By: Malak Essam **202000935**

Mazen Mohamed **202000687**

Youssef Ismail
202001864

Farah Amr
202001483

Table of Contents

Abstract

Post-weaning diarrhea is a significant health concern in pig farming, and the use of antimicrobials for disease control is facing challenges due to the emergence of antimicrobial resistance. Descriptive statistics, graphics, outlier detection, testing for normality and homoscedasticity, statistical inference, hypothesis testing, and linear modeling are used to analyze the data. In this study, we analyzed a dataset using the R programming language to investigate the effects of different treatments on average daily weight gain (ADWG) in piglets during the post-weaning period. The dataset included information on five treatment groups, including the use of zinc oxide (ZnO), nutraceuticals, and vaccination in different combinations. The study was conducted in pens with 16 piglets each, with a total of 128 piglets in each treatment group randomly allocated to 40 pens. The findings of this study provide insights into the potential effects of different treatments on piglet growth during the post-weaning period, which can contribute to the development of effective management strategies for post-weaning diarrhea in swine farms.

Introduction

Post-weaning diarrhea (PWD) is a significant issue in pig farming worldwide, causing increased mortality, weight loss, retarded growth, and higher treatment costs. Enterotoxigenic Escherichia coli is the primary cause of PWD, and the current control strategy involves the use of antimicrobials. However, the emergence of antimicrobial resistance in E. coli necessitates alternative control strategies such as dietary fiber and reduced crude protein levels. Additionally, the use of zinc oxide (ZnO) has shown positive effects in controlling the disease, but EU legislation may prohibit its use in 2022. Vaccination of piglets is another potential control strategy, and this study aims to compare the effects of vaccination to the addition of ZnO and nutraceuticals (e.g. fibers) to the feed. Five treatments were evaluated, and the study measured average daily weight gain (ADWG0021, ADWG2150, and ADWG0050) in different post-weaning periods. The study design involved pens of 16 piglets, with 128 piglets per treatment group. The data were randomized over the 5x8=40 pens. This project aims to conduct a statistical analysis of the data to evaluate the effectiveness of the different treatments in controlling PWD. The analysis will involve descriptive statistics, graphical representations, outlier detection, testing for normality/homoscedasticity, statistical inference, hypothesis testing, and linear modeling.

3. Methodology

3.1. Study design

This study is a randomized controlled trial conducted to investigate the effect of vaccination, zinc oxide, and nutraceuticals on post-weaning diarrhea (PWD) in pigs. The study involved five treatment groups (A-E) with a total of 640 piglets, each group consisting of 128 piglets. Piglets were housed in pens (cages) with 16 piglets in each pen. Treatment groups were randomly assigned to the pens.

3.2. Data collection

The study collected data on three outcomes: average daily weight gain (ADWG) between 0-21 days post-weaning (ADWG0021), ADWG between 21-50 days post-weaning (ADWG2150), and ADWG between 0-50 days post-weaning (ADWG0050). The data was collected by weighing all the piglets in a pen together at the beginning and end of each time and then calculating the ADWG. The study also collected data on the treatment groups coded as categorical variables (A-E).

3.3 Data analysis:

Data analysis was conducted using R statistical programming language. The data was first summarized using descriptive statistics, including means, medians, minimum and maximum values, and first and third quartiles for each outcome variable. Frequency tables were generated for the categorical variable, the treatment, and the sex group. Correlation coefficients were calculated between ADWG0021 and ADWG2150, as well as between ADWG0021 and ADWG0050.

Data visualization was also used to explore the data. Bar charts were generated to display the distribution of the categorical variable, the treatment group. Histograms were created to visualize the distribution of the continuous variables, ADWG2150 and ADWG0021. A scatterplot was generated to show the relationship between ADWG0050 and ADWG0021, with separate regression lines for males and females. Boxplots were also created to compare the distribution of ADWG0021 across the different treatment groups.

Outliers in the data were identified, but not removed. Normality and homoscedasticity were checked using two methods each for all variables mentioned above. Confidence intervals were calculated for the means of ADWG0021 per gender using 90%, 95%, and 99% confidence levels. Hypothesis testing was conducted to test the difference in ADWG0021 between males and females, as well as between the different treatment groups, using statistical hypothesis frameworks and post hoc testing. A linear regression was fitted to the data to interpret the regression coefficients for one of the hypotheses.

4. Results

4.0 Data reading

The data contains nine variables which are:

1. **Pen:** Pens containing 16 piglets in each pen.
2. **Treatment**: A, B, C, D, E.
3. **Feeder**: The food.
4. **Sex**: Male or Female.
5. **W0**: Pig's weight.
6. **P0**: Piglets of each pen.
7. **ADWG0021**: Between 0-21 days post-weaning.
8. **ADWG2150**: Between 21-50 days post-weaning.

9. **ADWG0050**: Between 0-50 days post-weaning.

4.1 Descriptive Statistics

1) We summarized our data and we found two categorical variables (Sex, Treatment). We calculated the mean, median, minimum, maximum, first and third quartile for each variable.

```
> summary(myData)
      Pen        Treatment     Feeder          Sex          W0              P0
 Min.   : 1.00   A:8       Min.   : 1.00   Female:20   Min.   : 87.50   Min.   :16
 1st Qu.:10.75   B:8       1st Qu.: 5.75   Male  :20   1st Qu.: 90.88   1st Qu.:16
 Median :20.50   C:8       Median :10.50               Median : 99.00   Median :16
 Mean   :20.50   D:8       Mean   :10.50               Mean   : 99.17   Mean   :16
 3rd Qu.:30.25   E:8       3rd Qu.:15.25               3rd Qu.:106.75   3rd Qu.:16
 Max.   :40.00             Max.   :20.00               Max.   :113.50   Max.   :16
    ADWG0021        ADWG2150        ADWG0050
 Min.   :102.7   Min.   :375.0   Min.   :275.6
 1st Qu.:129.5   1st Qu.:471.4   1st Qu.:325.7
 Median :144.3   Median :501.1   Median :349.7
 Mean   :143.1   Mean   :500.9   Mean   :350.6
 3rd Qu.:155.9   3rd Qu.:535.0   3rd Qu.:374.5
 Max.   :178.6   Max.   :608.1   Max.   :416.9
> str(myData)
'data.frame':   40 obs. of  9 variables:
 $ Pen      : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Treatment: Factor w/ 5 levels "A","B","C","D",..: 1 1 3 3 5 5 2 2 4 4 ...
 $ Feeder   : num  1 1 2 2 3 3 4 4 5 5 ...
 $ Sex      : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
 $ W0       : num  110 111 108 99 103 ...
 $ P0       : num  16 16 16 16 16 16 16 16 16 16 ...
 $ ADWG0021 : num  167 152 131 152 134 ...
 $ ADWG2150 : num  526 472 608 569 502 ...
 $ ADWG0050 : num  375 338 408 394 348 ...
```

2) Then we calculated the frequency table for the categorical variables.

```
> table(myData$Sex)

Female   Male
    20     20
> table(myData$Treatment)

A B C D E
8 8 8 8 8
```

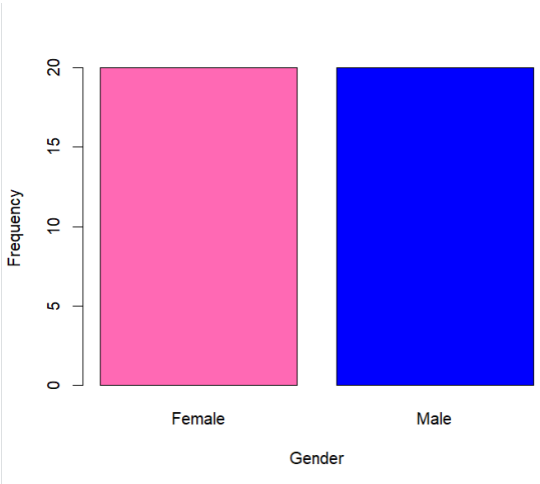
3) The correlation coefficient is the specific measure that quantifies the strength of the linear relationship between two variables in a correlation analysis. We calculated the correlation coefficient (ADWG0021 and ADWG2150) and (ADWG0021and ADWG0050) by using Pearson method. There is a positive relationship between (ADWG0021 and ADWG2150) and (ADWG0021and ADWG0050).

```r
#Correlation coefficient (ADWG0021 and ADWG2150)

#pearson
corrp1=cor(myData$ADWG0021,myData$ADWG2150 ,use="complete.obs",method="pearson")
corrp1
#0.2270356

#Correlation coefficient (ADWG0021and ADWG0050)

#pearson
corrp2=cor(myData$ADWG0021 , myData$ADWG0050 ,use="complete.obs", method = 'pearson')
corrp2
#0.4427165
#Positive realtionship
```

4.2 Graphical Analysis

- **Bar chart of a categorical variable for the gender (Sex parameter).**
  - We plotted a bar chart that provides a visual representation of the distribution of different categories in the sex variable which are male and female, allowing for easy comparison between the number of each one in our sample.
  - The x-axis corresponds to the variable categories, which are "Female" and "Male". The y-axis represents the frequency or count of each category in our sample.

```r
myData$Sex = factor(myData$Sex, levels=c(1,2), labels = c("Female","Male"))

barplot(table(myData$Sex), xlab="Gender",ylab="Frequency", col=c("hotpink","blue
```

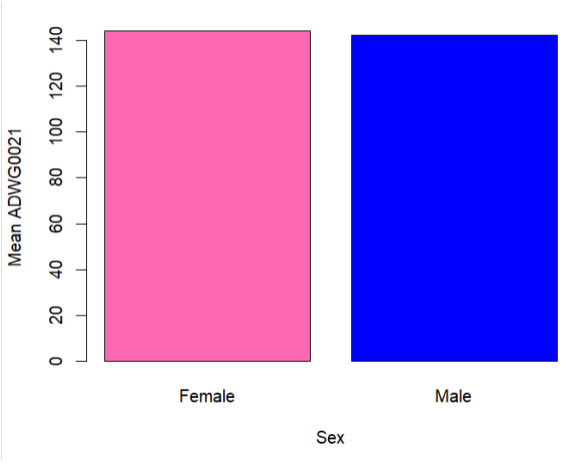


The graph showed: that number of Males and Females both are equal to 20.

- **Bar chart graph with mean ADWG0021 in males and females.**
  - We plotted a bar chart to easily compare the mean number of ADWG0021 (average daily weight gain in the first 21 days post-weaning) between males and females.
  - The tapply function is used to summarize data based on different groups or categories (males and females).

- The x-axis is the sex variable, and the y-axis is the mean ADWG0021 in each gender.

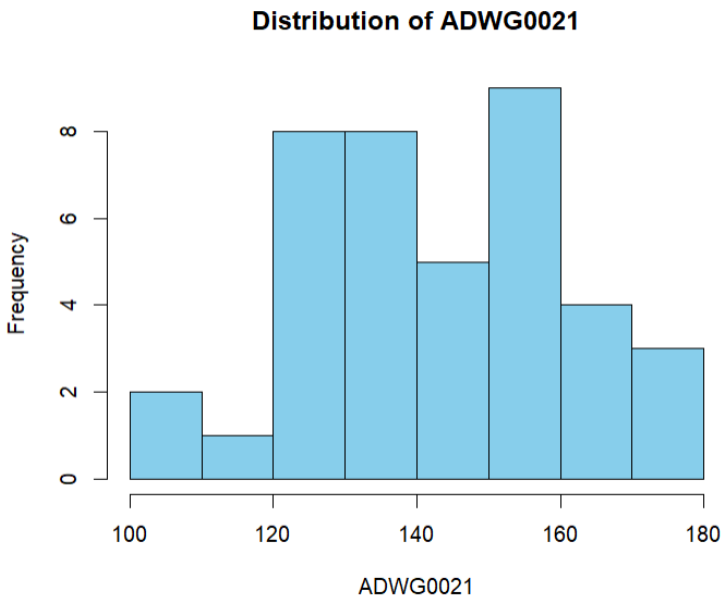
```
barplot(tapply(myData$ADWG0021,list(sex=myData$Sex),mean,na.rm=T),
        xlab="Sex",ylab="Mean ADWG0021",col=c("hotpink","blue"))
```

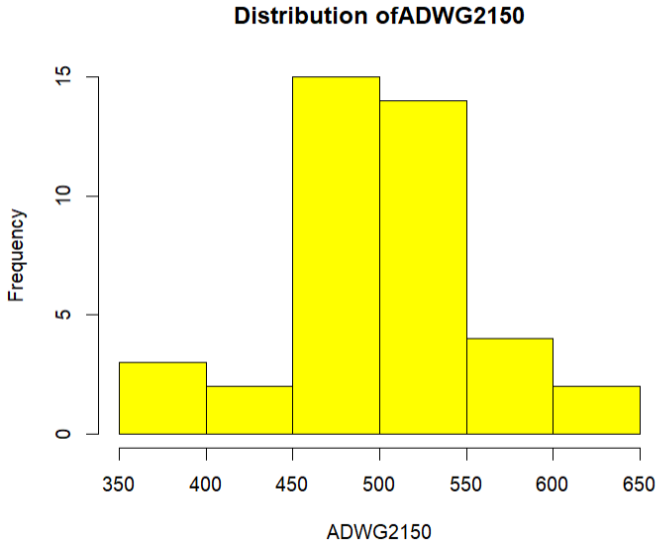


The graph showed: There is no significant difference in the average daily weight gain between male and female pigs in the period between 0- and 21-days post-weaning.

- **Histogram of a continuous variable: "ADWG2150" as well as "ADWG0021".**
  - We created a histogram for each continuous variable histogram to visualize the distribution of the ADWG0021 and ADWG2150.
  - The histogram provides an overview of the distribution, allowing you to observe patterns, central tendency, and variability in the data.
  - The histogram is also used to indicate the normality of the data. If the distribution was bell-shaped, this means that the data might be normal, if not then this data is not normal. (To formally test normality, we use Shapiro test).

```
hist(myData$ADWG2150,col="yellow", xlab="ADWG2150", main="Distribution ofADWG2150
hist(myData$ADWG0021,col="skyblue", xlab="ADWG0021", main="Distribution of ADWG00
```





Graph showed:

In ADWG2150:

- The x-axis represents the range of values for ADWG2150 (range is between 350 and 650), and the y-axis represents the frequency or count of observations within each bin.
- The data is symmetric which suggests that it is normal.
- There is a narrow distribution with a high peak which indicates that most of the piglets are gaining weight at a similar rate. This means most of the piglets gained 450-550 g per day in the period of day 21 and day 50 post-weaning.

In ADWG0021:

- There is a wide distribution with multiple peaks indicating more variability in weight gain across the piglets. The average daily weight gain is small in the first 21 days post-weaning.

- The x-axis represents the range of values for ADWG0021 (range is between 100 and 180), and the y-axis represents the frequency or count of observations within each bin.
- The data is symmetric which suggests that it is normal.

➢ **Scatterplot of 2 continuous variables ADWG0050 and ADWG0021 and add the regression lines for each gender.**

- First, we started by making data frame for each gender.

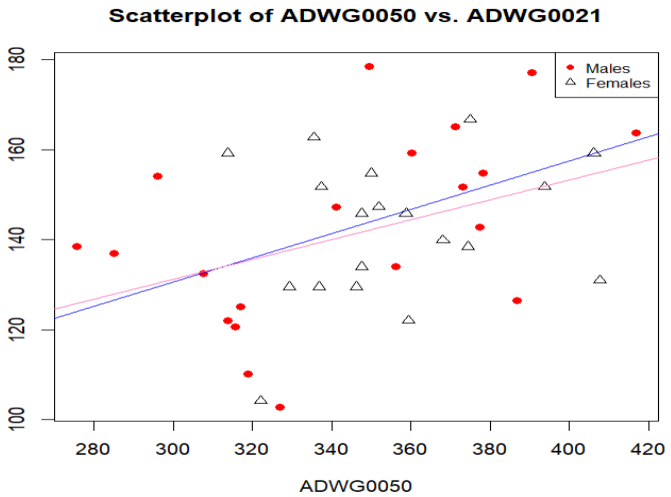
```
male_data <- myData[myData$Sex == "Male",]
female_data <- myData[myData$Sex == "Female", ]
```

- Then, we generated two scatterplots, one for males and one for females, with regression lines indicating the relationship between ADWG0050 and ADWG0021 for each gender.
- The x-axis is the ADWG in the first 50 days and the Y-axis is the ADWG in the first 21 days. The regression lines represent each gender.

```
plot(male_data$ADWG0050, male_data$ADWG0021, xlab = "ADWG0050", ylab = "ADWG0021"
     main = "Scatterplot of ADWG0050 vs. ADWG0021 for Males",pch = 16)
abline(lm(male_data$ADWG0021 ~ male_data$ADWG0050), col = "blue")

points(female_data$ADWG0050, female_data$ADWG0021, xlab = "ADWG0050", ylab = "ADW
       main = "Scatterplot of ADWG0050 vs. ADWG0021 for Females",pch = 16)
abline(lm(female_data$ADWG0021 ~ female_data$ADWG0050), col = "hotpink")
```

Graph showed:

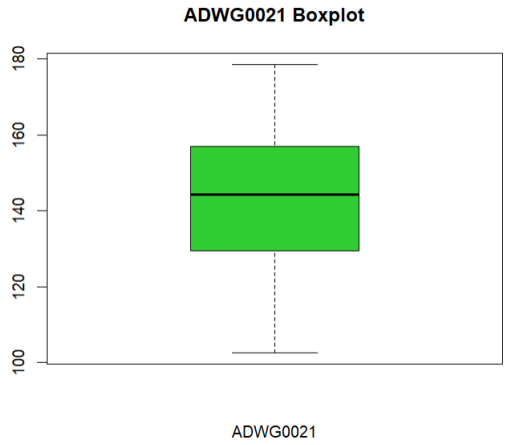
- There is no strong correlation or association between variables and the regression lines have a positive slope, it suggests a weak and inconsistent relationship where changes in one variable tend to be accompanied by corresponding but not highly reliable changes in the other variable. The wide scatter of data points implies that other factors, such as individual variation, environmental factors, or random fluctuations, are influencing the relationship, making it weak and inconsistent.



➢ **Boxplot of ADWG0021.**

- From boxplots we can visualize its distribution and conclude the median, quartiles, range, and variance of the average daily weight gain in the period between 0- and 21-days post-weaning.

```
boxplot(myData$ADWG0021,main="ADWG0021 Boxplot",xlab="ADWG0021",
```



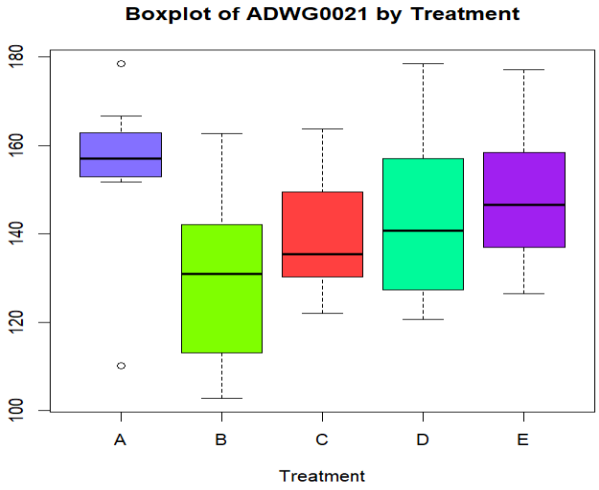
Graph showed:

- Interquartile Range (IQR): The box represents the interquartile range, which spans the middle 50% of the data. The lower edge of the box is around 130, and the upper edge is around 160. This means that the middle 50% of the values for ADWG0021 fall within this range.
- Whiskers: The lower whisker extends to approximately 100, which means that the data points below this value are considered within the range of the dataset. The upper whisker reaches approximately 180, indicating that the data points above this value are considered within the range.
- There are no outliers in the data.

➢ Separate boxplots per Treatment (as. factors)

- The separate boxplots for each Treatment group allow us to compare the distribution of ADWG0021 across different treatment conditions. Each boxplot represents the distribution of ADWG0021 within a specific Treatment group. By comparing the medians, the IQRs, and the whiskers of the boxplots, we can identify differences in the average daily weight gain among the Treatment groups.

```
boxplot(myData$ADWG0021 ~ as.factor(myData$Treatment), xlab = "Treatment", ylab = "ADWG0021",
        main = "Boxplot of ADWG0021 by Treatment",col=c("coral1","darkorchid1","mediumspringgreen","violetr
```



Graph showed:

- We can observe that Treatment groups B and D show noticeably higher daily weight gain compared to others. (because they have longer whiskers and they have taller boxes which show their high variability.
- Treatment C and Treatment E indicates a moderate spread of data points and average daily weight gain. This suggests a relatively consistent range of outcomes within these Treatment groups.
- Treatment A has a very small box, indicating a narrow IQR and potentially less variability in average daily weight gain. This suggests that there may be less diversity or a narrower range of outcomes within this Treatment group.
- There are two outliers in treatment A.

4.3 Outlier detection

We Explored the data by using boxplots for any existing outliers, identified them but we didn't remove them.

```
140  box0021 <- boxplot(ADWG0021~Treatment,data=myData, plot = FALSE)
141  box0021$out
142  box2150 <- boxplot(ADWG2150~Treatment,data=myData, plot = FALSE)
143  box2150$out
144  box0050 <- boxplot(ADWG0050~Treatment,data=myData, plot = FALSE)
145  box0050$out
```

We found two outliers in ADWG0021 per Treatment(A) with values 178.5714 and 110.1190 located outside the whiskers of the box plot

And another two in ADWG2150 and ADWG0050 per Treatment(B) with values 375 and 275.625 located outside the whiskers of the box plot.

4.4 Testing Normality, Homoscedasticity

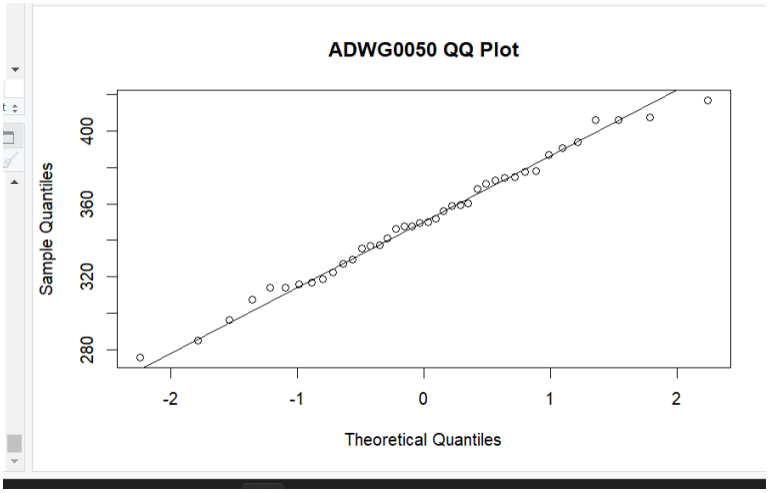
Normality Testing:

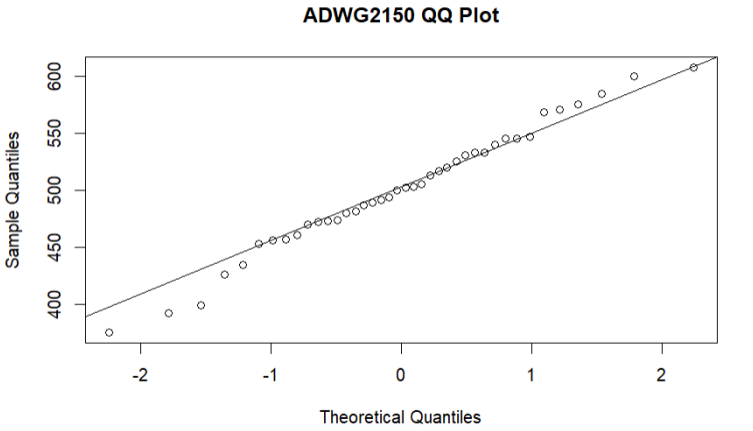
To assess the normality assumption, we performed a normality test on the data using the Shapiro-Wilk test, qq-plot. This test evaluates whether the data significantly deviate from a normal distribution. The null hypothesis (H0) assumes normality, while the alternative hypothesis (H1) suggests a departure from normality.

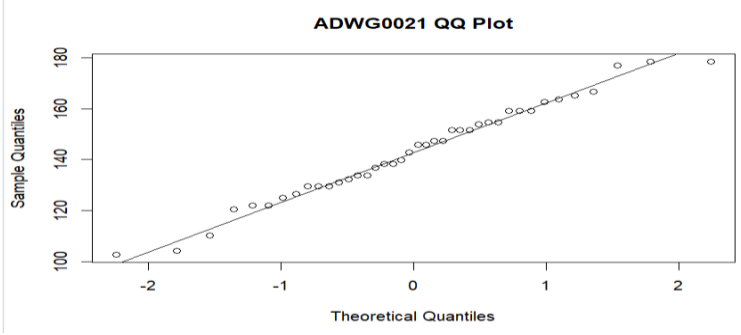
The results of the normality test indicated that the data were normally distributed as the p-value for ADWG0021, ADWG0050, and ADWG2150 are greater than the significance level of 0.05.

of p value= (0.7305,0.9086,0.9037)

which mean we don't have enough evidence to reject null hypothesis. It is worth noting that even if the normality assumption is violated, certain statistical tests are robust to moderate departures from normality, especially when the sample size is large. Additionally, visual examination of histograms, Q-Q plots, or other graphical methods can provide further insights into the distributional characteristics of the data.








```
> shapiro.test(myData$ADWG0021)

        Shapiro-Wilk normality test

data:  myData$ADWG0021
W = 0.98111, p-value = 0.7305

> shapiro.test(myData$ADWG0050)

        Shapiro-Wilk normality test

data:  myData$ADWG0050
W = 0.98657, p-value = 0.9086

> shapiro.test(myData$ADWG2150)

        Shapiro-Wilk normality test

data:  myData$ADWG2150
W = 0.98638, p-value = 0.9037

>
```

Homoscedasticity Testing:

To examine the assumption of homoscedasticity, we conducted Bartlett's test, a Levene’s test.

This test assesses whether the variances of the dependent variable are equal across different levels of the independent variables.

The null hypothesis (H0) assumes homoscedasticity, while the alternative hypothesis (H1) suggests heteroscedasticity.

The results of the homoscedasticity test revealed that the assumption of homoscedasticity was not violated as all the values of p-values are greater than the significance level of 0.05 (0.8498,0.09321,0.1251)

which means that we do not have sufficient evidence to reject the null hypothesis.

```
> leveneTest(myData$ADWG0021, myData$Treatment)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  4  0.2676 0.8968
      35
> leveneTest(myData$ADWG0050, myData$Treatment)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  4  1.4199 0.2478
      35
> leveneTest(myData$ADWG2150, myData$Treatment)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  4  1.3071 0.2864
      35
> bartlett.test(myData$ADWG0021, myData$Treatment)

        Bartlett test of homogeneity of variances

data:  myData$ADWG0021 and myData$Treatment
Bartlett's K-squared = 1.3676, df = 4, p-value = 0.8498

> bartlett.test(myData$ADWG0050, myData$Treatment)

        Bartlett test of homogeneity of variances

data:  myData$ADWG0050 and myData$Treatment
Bartlett's K-squared = 7.9558, df = 4, p-value = 0.09321

> bartlett.test(myData$ADWG2150, myData$Treatment)

        Bartlett test of homogeneity of variances

data:  myData$ADWG2150 and myData$Treatment
Bartlett's K-squared = 7.2119, df = 4, p-value = 0.1251
```

4.5 Statistics inference

We calculated the 90%, 95%, 99% confidence interval for the means of ADWG0021per each gender.

We started by calculating observations (n), mean, standard deviation (sd), standard error of the mean (sem) for each gender. Then, we calculated the Confidence Interval = Mean ± (t-value * Standard Error).

```
library(dplyr)

result <- myData %>%
  group_by(Sex) %>%
  summarize(
    n = n(),
    mean = mean(ADWG0021),
    sd = sd(ADWG0021),
    sem = sd / sqrt(n),
    ci90_lower = mean - qt(0.95, n - 1) * sem,
    ci90_upper = mean + qt(0.95, n - 1) * sem,
    ci95_lower = mean - qt(0.975, n - 1) * sem,
    ci95_upper = mean + qt(0.975, n - 1) * sem,
    ci99_lower = mean - qt(0.995, n - 1) * sem,
    ci99_upper = mean + qt(0.995, n - 1) * sem,
    interval_width90 = ci90_upper - ci90_lower,
    interval_width95 = ci95_upper - ci95_lower,
    interval_width99 = ci99_upper - ci99_lower

  )
```

Sex	n	mean	sd	sem	ci90_lower	ci90_upper	ci95_lower	ci95_upper	ci99_lower	ci99_upper
Female	20	144.0724	17.47262	3.906997	137.3167	150.8281	135.8950	152.2499	132.8948	155.2501
Male	20	142.1528	21.09261	4.716452	133.9974	150.3081	132.2811	152.0244	128.6593	155.6462

Then we calculated the width of each interval by subtracting the upper bound of each interval from the lower bound.

Value obtained were:

interval_width90	interval_width95	interval_width99
13.51143	16.35488	22.35533
16.31074	19.74329	26.98692

• So, to describe those inferences?

- we can describe the inferences by stating that we are 90%, 95%, and 99% confident that the true mean ADWG0021 for males and females falls within the calculated confidence intervals.
- We noticed that by increasing the confidence level, the width of the confidence interval also increases, indicating increased uncertainty but higher confidence in capturing the true mean.

4.6 Hypothesis testing

4.6.1. We hypothesis that ADWG0021is different between male vs female. Assuming normality and homoscedasticity, can you test this hypothesis using statistical hypothesis framework.

To test the hypothesis that ADWG0021 is different between males and females, we Formulated the Null Hypothesis that: There is no difference in ADWG0021 between males and females and the Alternative Hypothesis: There is a difference in ADWG0021 between males and females. we used a Two-sample independent t-test which compares the means of two independent groups (males and females), we Assumed Normality and equal variances between the groups. the reason why we used the t-test is That t-test is commonly used to compare the means of two groups when the data is assumed to be normally distributed and the variances are assumed to be equal (results of Testing for normality/ homoscedasticity) Specifically

for our hypothesis of comparing ADWG0021 between males and females, assuming normality and homoscedasticity. our code performs an independent samples t-test to compare the mean ADWG0021 between males and females based on the "Sex" variable in our data.

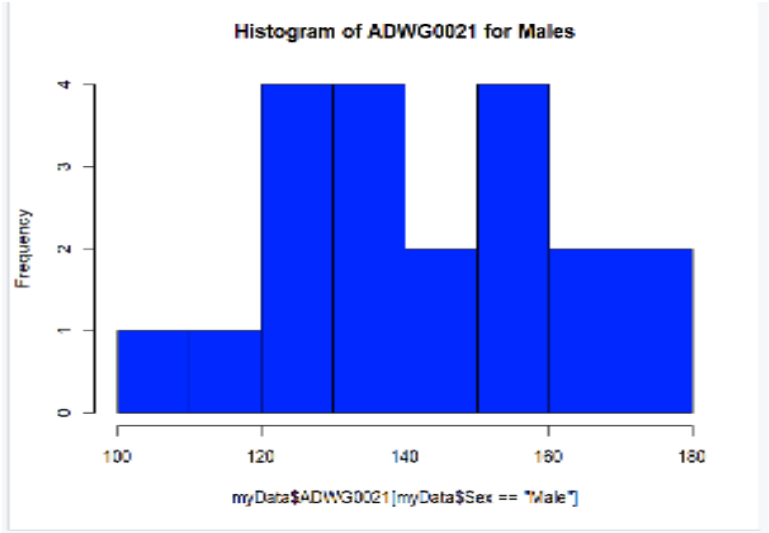
Results: the p-value = 0.7557 which is greater than the alpha of value 0.05, which means that we don't have enough evidence to reject the null hypothesis that There is no difference in ADWG0021 between males and females

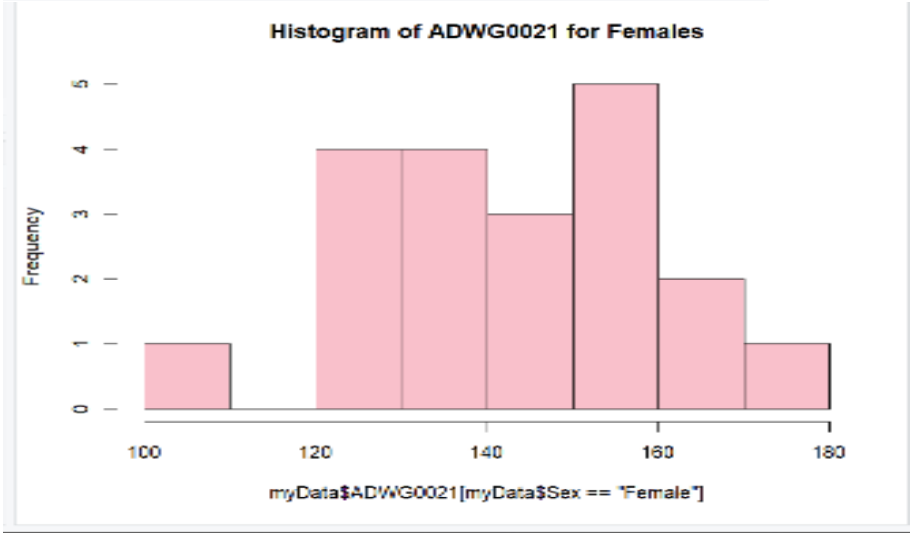
```
> t_test_result=t.test(ADWG0021 ~ Sex, data = myData)
> print(t_test_result)

        Welch Two Sample t-test

data:  ADWG0021 by Sex
t = 0.31344, df = 36.728, p-value = 0.7557
alternative hypothesis: true difference in means between group Female and group Male is not equal
to 0
95 percent confidence interval:
 -10.49288  14.33217
sample estimates:
mean in group Female   mean in group Male
            144.0724             142.1528

>
```

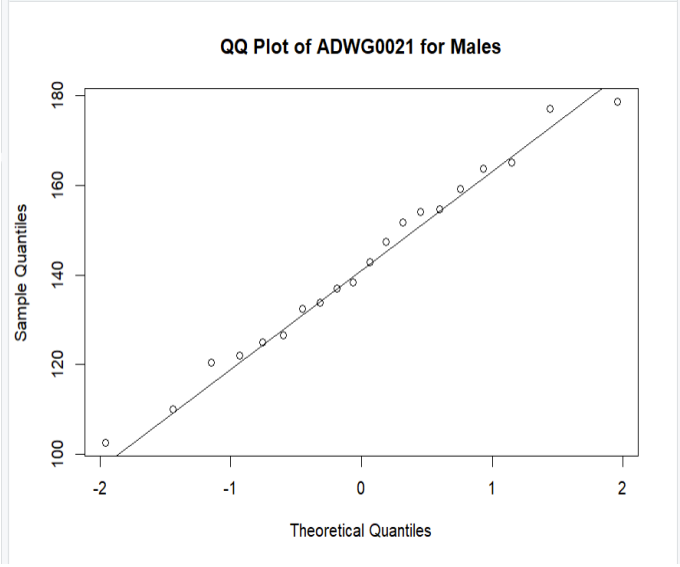
4.6.1.1. To assess whether the assumptions for the t-test have been met we evaluated the following assumptions:

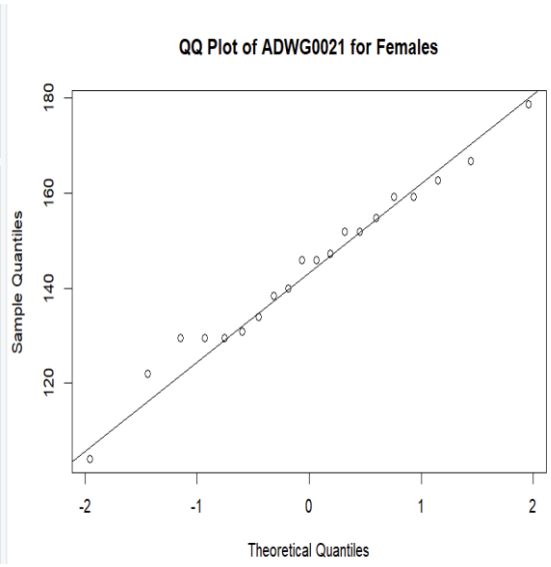
Normality:

The t-test assumes that the data within each group (males and females) are normally distributed. We assessed normality visually using a histogram and a QQ plot, as well as statistically using the Shapiro-Wilk test.









Results:

The p-value from the Shapiro-Wilk test is greater than the significance level, it indicates that there is not enough evidence to conclude that the data significantly deviate from a normal distribution.

```
        Shapiro-Wilk normality test

data:  myData$ADWG0021[myData$Sex == "Male"]
W = 0.98144, p-value = 0.9513

> shapiro.test(myData$ADWG0021[myData$Sex == "Female"])

        Shapiro-Wilk normality test

data:  myData$ADWG0021[myData$Sex == "Female"]
W = 0.98432, p-value = 0.9771
```

Equality of variances: The t-test assumes that the variances of the two groups are equal. We checked this assumption using Levene's test or the F-test for equality of variances.

Results:

Based on these results and a significant level of alpha= 0.05, there is no significant evidence to suggest differences in variances between the groups or a significant difference in the ratio of variances. Therefore, you can proceed with the assumption of equal variances in your analysis.

```
> leveneTest(ADWG0021 ~ Sex, data = myData)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  1.0594 0.3099
      38
> var.test(ADWG0021 ~ Sex, data = myData)

        F test to compare two variances

data:  ADWG0021 by Sex
F = 0.68621, num df = 19, denom df = 19, p-value = 0.4193
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2716092 1.7336692
sample estimates:
ratio of variances
         0.6862074
```

4.6.2. We hypothesis that ADWG0021is “different” in the group receiving Treatment A (normal feed + ZnO) compared to the Treatment B (normal feed + nutraceuticals). Can you test this hypothesis assuming heteroscedasticity.

```r
#We hypothesis that ADWG0021is “different” in the group receiving
#Treatment A (normalfeed + ZnO) compared to the Treatment B (normal feed + nutraceuticals)
#Can you test thishypothesis assuming heteroscedasticity.
table(PWD$Treatment)
A=(PWD$ADWG0021[PWD$Treatment=="A"])
B=(PWD$ADWG0021[PWD$Treatment=="B"])
#Paired t-test:
  t.test(A, B, paired= T, var.equal = F)
  # mean difference 24.59077
  #mean difference is not equal to 0

#We use Bartlett's test of homoscedasticity:
bartlett.test(list(A[A == A], B[B == B]))
# p-value = 0.955
# our data is not heteroscedastic (they have the same variance).

##############################################
```

Result:

Paired t-test

-mean difference 24.59077

- mean difference is not equal to 0

Bartlett test of homogeneity of variances

p-value = 0.955

4.6.2.1. Assess the previous test assumption.

To test the homogeneity of variance, we use Bartlett's test of homoscedasticity. In Bartlett's test the null hypothesis (HO) means there is no difference in variance, the alternative (H1) means there is differences in variance (Heteroscedastic). From the test results we can see that p-value equals to 0.955 which is larger than the significant level alpha (0.05), so we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis. This concludes that our data is not heteroscedastic (they have the same variance).

4.6.3. We hypothesis that ADWG0021 is different between the different Treatments.

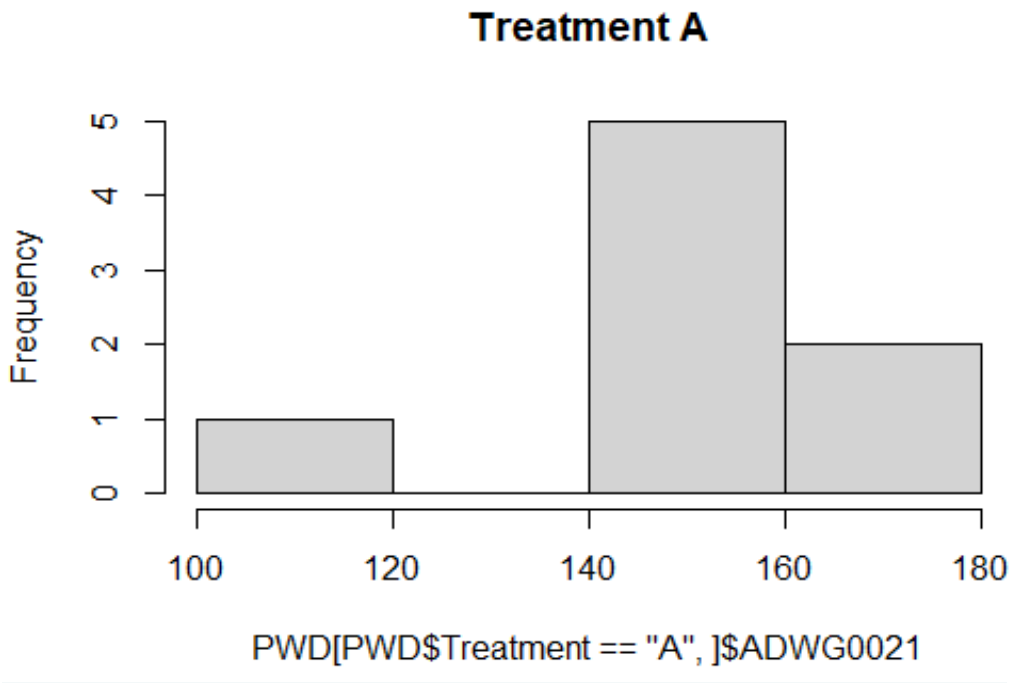
The statistical question: Is the mean difference between the different Treatments in ADWG0021? We will test normality (Histogram, QQ plot, Shapiro test).

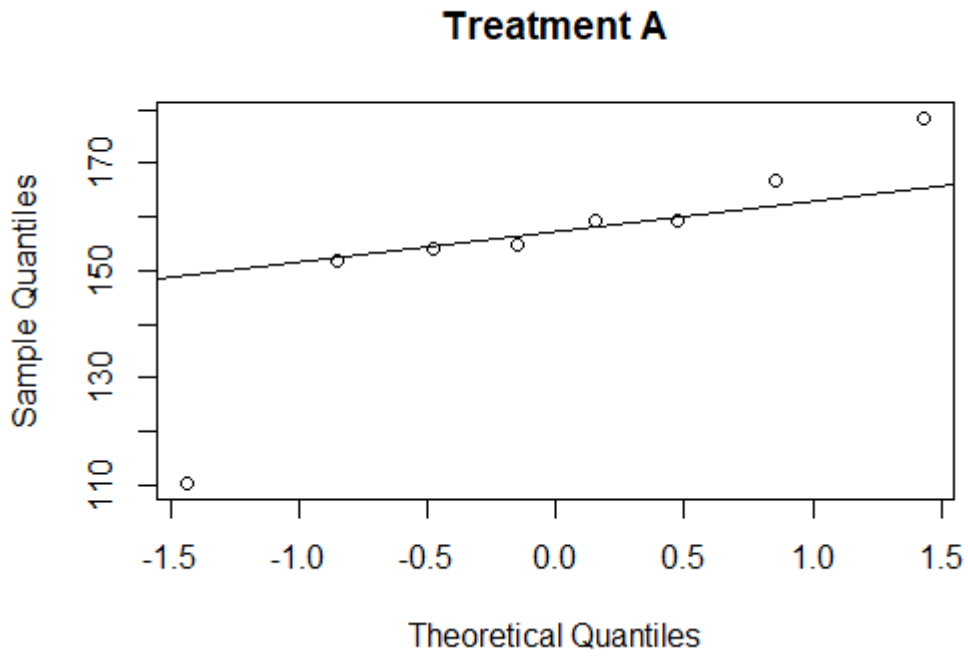
Null hypothesis: the mean of the treatment groups is equal.

Alternative hypothesis: the mean of the treatment groups is not equal.

Results:

Histogram and QQ plot show that only treatment A is not normally distributed.





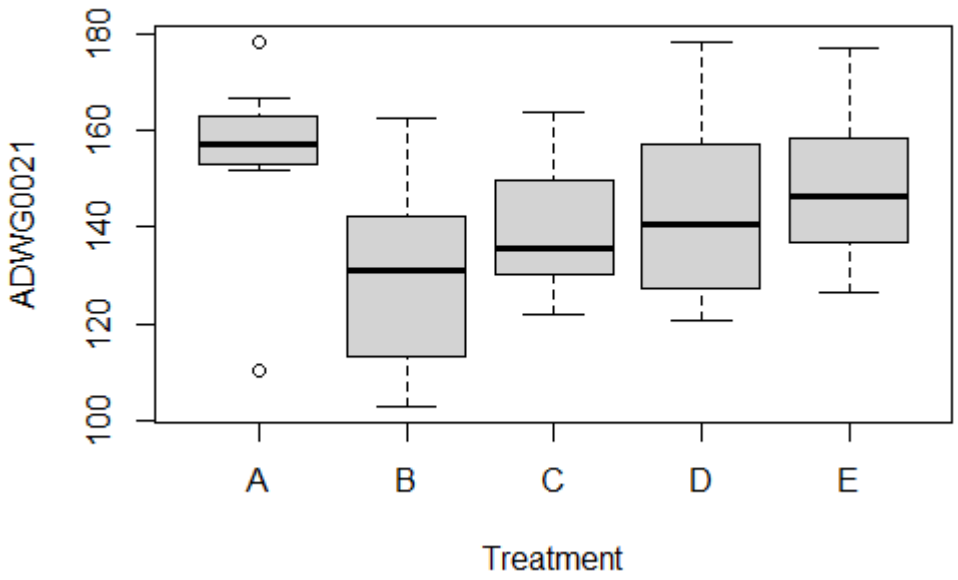
And in Shapiro test the treatment A is also not normal because the p-value is 0.0395 which is less than 0.05 so we have enough evidence to reject the null hypothesis. And the other treatments are normal with a p-value greater than 0.05.

```
        Shapiro-Wilk normality test

data:  PWD[PWD$Treatment == "A", ]$ADWG0021
W = 0.81314, p-value = 0.0395
```

Then we will use Kruskal Walis test (in case that the normality can't be assumed) because we have a data that is not normal and its p-value = 0.09972, so we don't have enough evidence to reject the null hypothesis which means that there is no difference between the treatments in ADWG0021. (no sample dominates another sample)

```
        Kruskal-Wallis rank sum test

data:  ADWG0021 by Treatment
Kruskal-Wallis chi-squared = 7.7864, df = 4, p-value = 0.09972
```

We will check the homoscedasticity by using (Boxplot and Levene test)




```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  4  0.2676 0.8968
      35
> |
```

In levene test (If normality can be assumed) gives a p-value 0.8968 which is greater than 0.05 so we do not have enough evidence to reject the null hypothesis so the data is homo variance.

```
> AnovaModel=aov(ADWG0021~Treatment, data= myData)
> AnovaModel
Call:
   aov(formula = ADWG0021 ~ Treatment, data = myData)

Terms:
                Treatment Residuals
Sum of Squares   2770.996 11519.480
Deg. of Freedom         4        35

Residual standard error: 18.14189
Estimated effects may be unbalanced
```

We will (assume normality and homoscedasticity). In ANOVA, the goal is to determine whether there are statistically significant differences between the means of two or more groups. ADWG0021 is the dependent variable, treatment is the independent variable. The ~ symbol is used to specify the relationship between the dependent and independent variables and store the results in the ANOVA Model object. Residuals in ANOVA are the differences between the dependent variable and the predicted values of it.

```
> summary(AnovaModel)
            Df Sum Sq Mean Sq F value Pr(>F)
Treatment    4   2771   692.7   2.105  0.101
Residuals   35  11519   329.1
```

The summary includes several statistics, such as F-statistic and p-value, which indicate whether the predictor variable have a significant effect on the dependent variable. The summary table includes the sum of squared errors and the mean squares, which are used to calculate the F-statistic. Treatment is statistically not significant and has F-value=2.105 and p-value=0.101 which is greater than 0.05 so we do not have enough evidence to reject the null hypothesis and ADWG0021 is not different between the different Treatments.

```
> report(AnovaModel)
The ANOVA (formula: ADWG0021 ~ Treatment) suggests that:

  - The main effect of Treatment is statistically not significant and large (F(4, 35) = 2.10, p
= 0.101; Eta2 = 0.19, 95% CI [0.00, 1.00])
```

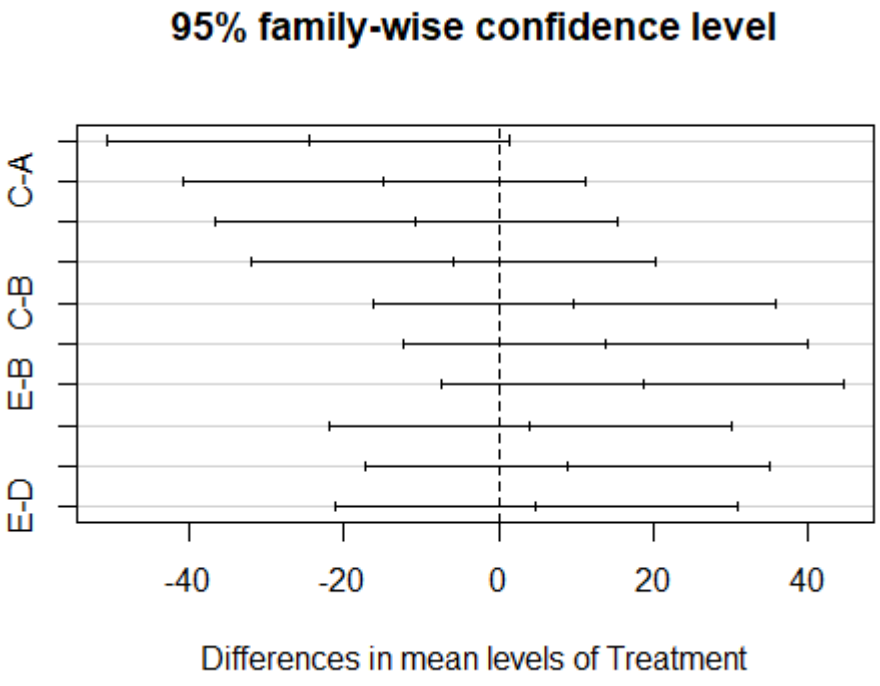
The report provides a detailed summary of the results.

```
> coef(AnovaModel)
(Intercept)  TreatmentB  TreatmentC  TreatmentD  TreatmentE
 154.303075  -24.590774  -14.794147  -10.701885   -5.865575
> |
```

It is used to extract the estimated coefficients of the independent variable from an ANOVA model, which can be used to interpret the effects of the independent variable.

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

Fit: aov(formula = ADWG0021 ~ Treatment, data = myData)

$Treatment
          diff        lwr       upr     p adj
B-A -24.590774 -50.670281  1.488734 0.0724945
C-A -14.794147 -40.873654 11.285361 0.4884256
D-A -10.701885 -36.781392 15.377622 0.7626339
E-A  -5.865575 -31.945083 20.213932 0.9661386
C-B   9.796627 -16.282880 35.876134 0.8154751
D-B  13.888889 -12.190619 39.968396 0.5499153
E-B  18.725198  -7.354309 44.804706 0.2580810
D-C   4.092262 -21.987246 30.171769 0.9910644
E-C   8.928571 -17.150936 35.008079 0.8605295
E-D   4.836310 -21.243198 30.915817 0.9832666

> |
```

The results of post hoc shows the difference of means between treatments with each other. All the treatments with each other gives a p adjusted value greater than 0.05 so we do not have enough evidence to reject the null hypothesis so this means that ADWG0021 is not different between the treatments.



The zero value lies in the middle of confidence interval in Tukey plot, so we won't reject the null hypothesis.

```
data: x and group
Kruskal-Wallis chi-squared = 7.7864, df = 4, p-value = 0.1


                         Comparison of x by group
                               (Bonferroni)
Col Mean-|
Row Mean |          A          B          C          D
---------+--------------------------------------------
       B |   2.621996
         |     0.0437
         |
       C |   1.808642  -0.813354
         |     0.3525     1.0000
         |
       D |   1.444773  -1.177223  -0.363868
         |     0.7426     1.0000     1.0000
         |
       E |   0.866864  -1.755132  -0.941778  -0.577909
         |     1.0000     0.3962     1.0000     1.0000

alpha = 0.05
Reject Ho if p <= alpha/2
>
```

This is a linear hypothesis between each pair of treatments. B-A treatments have a p-value=0.0437 which is smaller than 0.05 so we have enough evidence to reject the null hypothesis and the data is not normal and B-A treatments is different in ADWG0021.

4.7 Linear model

4.7.1 Fit a linear regression to the data and interpret the regression coefficient.

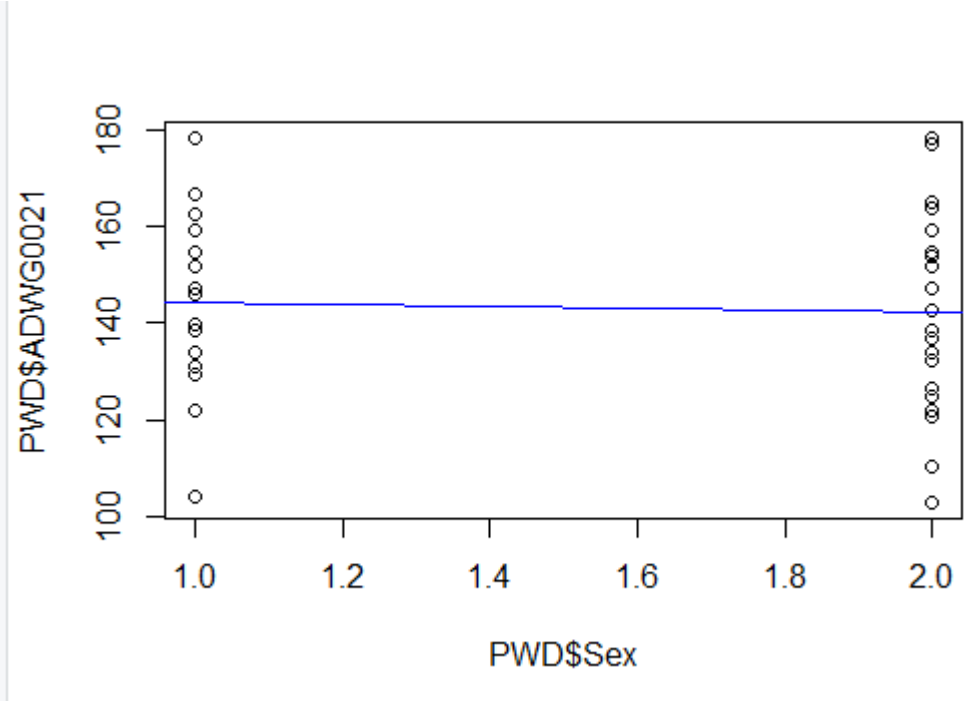
Null hypothesis: the slope equals zero which means y cannot be predicted by x.
Alternative hypothesis: the slope is not equal zero which means y can be predicted by x.
The ADWG0021 is continuous with sex which is categorical and we convert the sex to numeric we will assume that male=2, female=1, then make a plot and get intercept and slope.

```
Call:
lm(formula = ADWG0021 ~ Sex, data = PWD)


Coefficients:
(Intercept)          Sex
     145.99        -1.92
```

Intercept: The predicted value of Y when X equals zero.
Slope: Is the rate of change in y as x changes by a unit.
Intercept = 145.99 and slope = -1.92.

```
> summary(regression)

Call:
lm(formula = ADWG0021 ~ Sex, data = PWD)

Residuals:
    Min      1Q  Median      3Q     Max
-39.906 -14.608   1.233  13.245  36.419

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  145.992      9.684  15.076   <2e-16 ***
Sex           -1.920      6.125  -0.313    0.756
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.37 on 38 degrees of freedom
Multiple R-squared:  0.002579,  Adjusted R-squared:  -0.02367
F-statistic: 0.09824 on 1 and 38 DF,  p-value: 0.7557
```

The min and max have close values which is good and the median equals 1 which is close to zero so it is good. Here we measure the variance in the dependent variable (ADWG0021) (response) that can be predicted by the independent variable (Sex) (Explanatory) that means we have only one explanatory variable so this is a simple linear regression. We will use the multiple R-squared to know the percentage of Y being able to predict X, which equals 0.2579%.
The p–value=0.7557 which is greater than the significance level 0.05 so we do not have enough evidence to reject the null hypothesis (not significant), so Y can't be predicted by X.



There is no linear relationship between the explanatory variable (sex) with the response variable (ADWG0021) the regression line is horizontal seems to be zero so y (ADWG0021) can't be predicted by x (sex).

4.7.2 Calculate and interpret a 95% confidence interval of the regression slope.

First, we extract the coefficients (estimated regression parameters) from the linear regression model

regression.

gender_coefficient represents the estimated change in the response variable (ADWG0021) for a one-unit change in the Gender variable.

```
# Interpretation of regression coefficient
coef_lm <- coef(regression)
gender_coefficient <- coef_lm[2]
```

Then we calculate the 95% confidence interval for the coefficients of the linear regression model using the confint() function.

```
> conf_interval <- confint(lm_model, level = 0.95)
> conf_interval
                2.5 %    97.5 %
(Intercept) 135.30542 152.83942
SexMale     -14.31805  10.47877
```

Interpreting the confidence interval**,** the 95% confidence interval for the intercept is estimated to be between 135.30542 and 152.83942. This means that if the gender variable (SexMale) is held constant, we can be 95% confident that the true average value of ADWG0021 falls within this range.

The 95% confidence interval for the coefficient of SexMale is estimated to be between -14.31805 and 10.47877. This indicates that, with a 95% level of confidence, the true effect of being male (compared to female) on the average value of ADWG0021 is expected to fall within this range.

However, since the confidence interval for the coefficient of SexMale includes zero (suggesting that the effect of being male on ADWG0021 could be zero or negligible), we cannot conclude with 95% confidence that there is a statistically significant difference in the average value of ADWG0021 between males and females.

4.7.3 Estimating the average ADWG0021 changes with changing the gender from Female to Male.

```
# Estimating the average ADWG0021 change with changing gender from 1 to 2
gender_change <- 2 - 1
average_change <- gender_change * gender_coefficient
cat("Estimating the Average ADWG0021 Change for Changing Gender from 1 to 2:", rou
```

The coefficient -1.919643 indicates the estimated change in the average value of the response variable (ADWG0021) when the predictor variable (Sex) changes from Male to Female.

the result SexMale = -1.919643 means that, on average, there is an estimated decrease of approximately 1.92 units in the value of ADWG0021 when changing the gender from "Female" to "Male." This suggests that, compared to females, males have a lower average value of ADWG0021.

5. Conclusion

There is a positive relationship between (ADWG0021 and ADWG2150) and (ADWG0021 and ADWG0050). We found two outliers in Treatment A and Treatment B by using a box plot. We calculated the 90%, 95%, and 99% confidence interval for the means of ADWG0021 per gender and we noticed that by increasing the confidence level, the width of the confidence interval also increases, indicating increased uncertainty but higher confidence in capturing the true mean. Then we tested the hypothesis that

ADWG0021 is different between males and females and we noticed that there is no difference in ADWG0021 between males and females. We also tested the hypothesis that ADWG0021 is different in the group receiving Treatment A (normal feed + ZnO) compared to Treatment B (normal feed + nutraceuticals) and we noticed that the data is homoscedasticity. And we hypothesize that ADWG0021 is different between the different Treatments and we noticed that the data is homoscedasticity. Then we fit a linear regression to the data and we noticed that Y can't be predicted by X. We calculated the 95% confidence interval of the regression slope and the confidence interval for the coefficient of Sex(Male) includes zero so we cannot conclude with 95% confidence that there is a statistically significant difference in the average value of ADWG0021 between males and females.