STA 106 Project I Group 7

By: Alexander Derango, John Toland, Jiawei Zhu,

Valerie Whitfield, Devin Sidhu, Yanyu Zhu

Instructor: Maxime Guiffo Pouokam

2025-10-25

Figure 1: Picture of weight loss

1: Introduction

For this report, the dataset used contains the outcomes of an observational study which documented the effects that three different diet plans had on the weight loss of 76 subjects, with the plans labelled Diet A, Diet B, and Diet C. The dataset contains two columns; the categorical variable of which diet plan each subject was on and the numerical variable of how much weight they lost (in pounds) after six months on their respective diets. For this dataset, a positive number indicates the subject lost that amount of weight, while a negative number indicates that they gained that amount. For this dataset, the response variable is the amount of weight lost by subject, as that is the primary interest of the study, while the explanatory variable is the specific diet plan the subjects are on, as it is a possible explanation for the variability within the response variable.

The first interest of this report is to determine the best diet for losing weight, if there is any at all. To do this, a

Single Factor ANOVA test with a 5% significance level will be implemented, in order to determine if the averages of weight lost by diet plan are equal, or if at least one of the group means is significantly different from the others.

Lastly, if it is determined that at least one of the group means differs from the others, pairwise confidence interval tests will be conducted by comparing each possible pair of the diet plans with each other. This will help give a better understanding of the differences between the average weight loss among two groups, and show comparatively which diet plans are better or worse than others, and by how much. This can help determine the best diet plan for losing weight. Along with this, a factor effect model will be utilized to determine the contribution that each diet plan's average has on the sample mean as a whole, regardless of group.

2: Summary of the Data

Table 1: Summary Data Values

Diet	Sample.Size	Mean	Standard.Deviation
A	24	3.3000	2.2401
B	25	3.2680	2.4625
C	27	5.2333	2.2477

This table summarizes the observed data that was collected from all the 3 diet factor levels, Diet A, Diet B, and Diet C. This table includes the means, standard deviations, and sample size for each diet group. For Diet A, the sample size was 24 participants. The average weight loss was 3.3 pounds, and the standard deviation was 2.2401. For Diet B, the sample size was 25 participants. The average weight loss was 3.268 pounds and the standard deviation was 2.4625. For Diet C, the sample size was a little larger at 27 participants. The average weight loss was higher at 5.2333 pounds with a standard deviation of 2.2477. We can see that the sample sizes are very similar for all 3 diets, as well as the standard deviations. The average weight loss was higher in Diet C than in Diets A and B. Since the sample sizes are similar across the diet groups, we are able to compare the groups fairly.

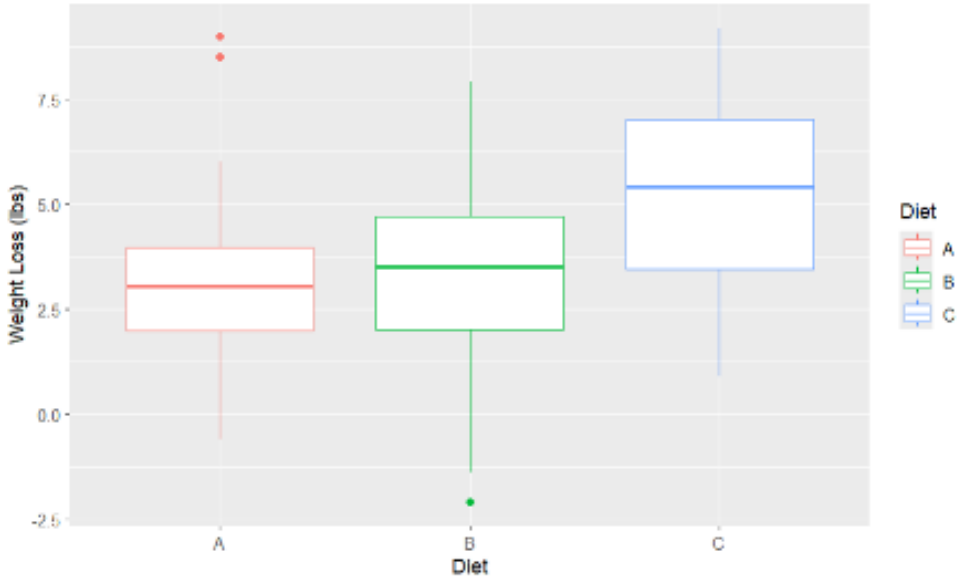

Figure 2: Boxplot of Weight Loss by Diet Groups

This boxplot above displays the variability in weight loss across the three diet groups. For Diet A, the median weight loss is approximately 3 lbs, with the lower quartile around 2 lbs and the upper quartile around 4 lbs. The overall spread is moderate, and there are two potential outliers, which are located around 9 lbs. This indicates that a few participants experienced unusually large weight losses. For Diet B, the median weight loss is slightly higher, around 3.5 lbs, with quartiles around 2 lbs and 4.75 lbs. This group contains one potential outlier at around -2 lbs. This suggests one participant actually gained a small amount of weight in this dieting process. The whiskers for Diet B extend slightly farther than in the other diets. This may indicate greater variability within this group. For Diet C, the median weight loss is the highest at around 5.5 lbs, with quartiles between 3.5 lbs and 7 lbs. There are no apparent outliers, and the data are relatively symmetrically distributed. This suggests consistent weight loss among participants following this diet. From the histogram, Diet B exhibits the greatest spread of all the three diets due to its longer whiskers and single negative outlier. Diet C and Diet A on the other hand appear to be more consistent due to their shorter whisker sizes and smaller inter quartile ranges. Across all of the three diet groups, most of the values are positive, which indicates that nearly all of the participants experienced at least some weight loss regardless of their assigned diet.

Figure 3: Histogram of Weight Loss by Diet Group

These three histograms above display the distribution of weight loss (in pounds) for each of the three diet groups. In the histogram, Groups A and B are both centered around approximately 3 pounds of weight loss. On the other hand, Group C is shifted to the right compared to A and B, with its distribution centered around 5 pounds. This may suggest that participants in this group generally experienced greater weight loss than the other diet groups. From the histograms, all three of the distributions appear roughly normal. However, it seems Groups B and C show a very slight left skew. In context, this would indicate that a few individuals lost slightly less weight than the group average. Each diet group has a similar number of participants, which is roughly 25 participants per diet group. This allows for a fair comparison between their distributions since the sample size is relatively equal. Overall, these histograms suggest that while Diets A and B yield solid results, Diet C appears to be the most effective in promoting weight loss.

3: Diagnostics

To find out if there were any outliers, we first fitted a model and added a column to our dataset for the estimated residuals. We then calculated the semi-studentized residuals from the residuals divided by the square root of our MSE, and compared them to the $t_{1-\alpha/(2 \cdot nt)}$ value with alpha = 0.05. By our values, we found no outliers according to that method.

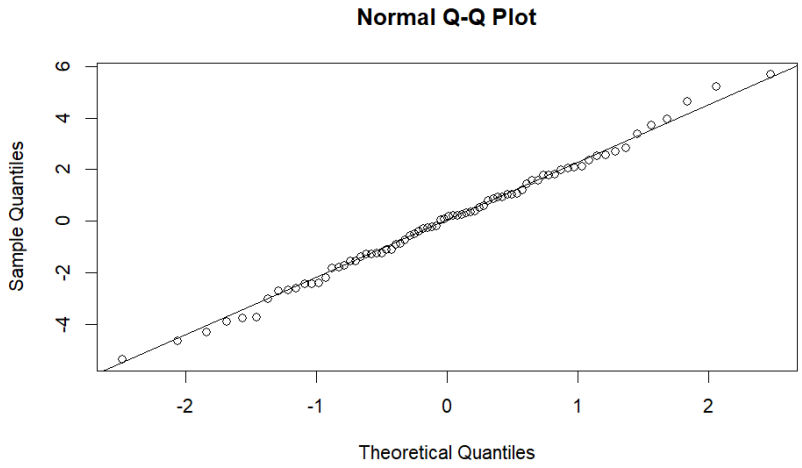

Figure 4: Normal Q-Q Plot

We created a normal Q-Q plot from the data. This plot has most of the points fairly close to the line. We then performed a Shapiro-Wilks test to test for whether the errors were normally distributed, and it returned a p-value of 0.9921. At our significance level of 0.05 (or any other reasonable level), we do not reject the null hypothesis of normally distributed errors and conclude that the errors are normally distributed. Additionally, we plotted errors vs group mean to verify whether the errors had roughly equal variance. Since they are distributed in similar amounts above and below the mean error of 0, the errors look to have roughly equal variance. A box plot of the errors by diet group gives a similar impression.

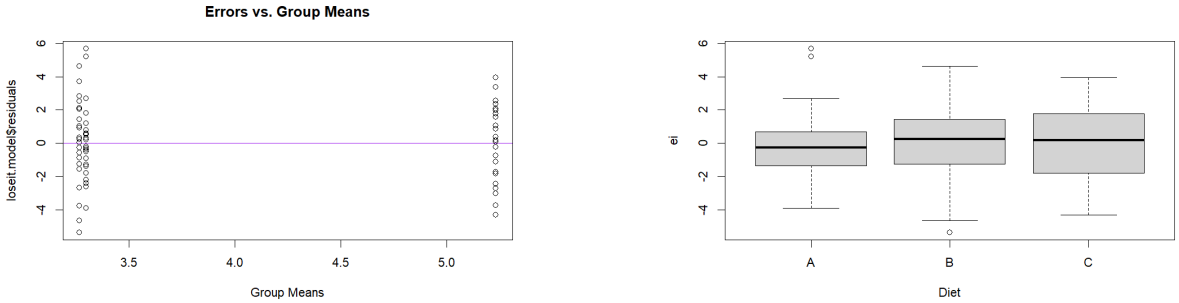

Figure 5: Errors vs. Group Means (left) and Boxplot of Errors by Diet (right)

To further test for homoscedasticity, we performed a Brown-Forsythe Test. We got a p-value of 0.694648, which at any reasonable alpha level we do not reject the null hypothesis and conclude that the variance of the errors are equal.

4: Analysis

Model Fit

Our goal for this analysis is to compare the group means, and establish whether there is a significant statistical difference between the different diet groups and their effects on weight loss. We will use the group mean model which is depicted below:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

We will be comparing the averages between diet A, diet B, and diet C, and testing to see if at least one is significantly different from the others. In this model μ represents the average weight change of an individual, having $i = A, B, C$, which are the three different diet groups. Finally, ϵ_j represents the error of the jth observation within group i. These averages will be estimated using ANOVA.

Hypothesis Testing

In our hypothesis tests, we will be looking at if there is a significant difference in the average weight loss between groups. Our null hypothesis, H_0, is that each group mean is equal, between the three groups, diet A, diet B, and diet C, meaning no difference between the groups. This can be depicted below as,

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

Alternatively, what is needed to reject the null hypothesis is for at least one of the group means differing from the other groups. Therefore we would have an alternative hypothesis depicted as,

H_A : At least one of the means of groups diet A, diet B, and diet C, differs from the other groups.

To come to such conclusions on whether or not to reject the null hypothesis, we must calculate the test statistic, and find the corresponding p-value. We can find these values using ANOVA tables, which computes the MSA, and MSE, which can be used to find the test statistic. The formula for finding the test statistic is MSA/MSE, and the greater its value, the less likely the null hypothesis is true. A test statistic of exactly one would infer that the group means are equal and there is no difference. We will also find the corresponding p-value, which when compared to a given confidence level, will determine for us whether or not to reject the null hypothesis. The output for our ANOVA table is the following:

Table 2: ANOVA TABLE

term	df	sumsq	meansq	statistic	p.value
Diet	2	66.1830	33.0915	6.1537	0.0034
Residuals	73	392.5544	5.3775	NA	NA

The test-statistic comes out to a value of 6.1537, and the corresponding p-value comes out to 0.0034. The p-value represents the probability of observing our data given that the null hypothesis is true, meaning that the probability is about %0.0034. Given a confidence level, we can compare our p-value to such an alpha value, which would determine whether or not to reject the null hypothesis.

Confidence Intervals

We will be computing confidence intervals of 95% comparing the group means, by running pairwise confidence intervals. The confidence intervals will be comparing the following:

$$\mu_A - \mu_B, \quad \mu_B - \mu_C, \quad \mu_A - \mu_C$$

We must compute the multiples used for the Confidence Intervals. Because the confidence interval is pairwise, we will be computing the Tukey and Bonferroni multiples. After computing both, we will take the smaller of the two values, and apply that as our multiple. The computed multiples are calculated as below

x	value
Tukey	2.3924
Bonferroni	2.4504

The smaller observed multiple is the given value by Tukey, thus we will be applying Tukey to the multiple, with a value of 2.3924, and will compute the three pairwise confidence intervals.

$$\mu_i - \mu_{i'} = (\bar{Y}_i - \bar{Y}_{i'}) \pm T \times \sqrt{\text{MSE} \times ((1/n_i) + (1/n_{i'}))} \text{Where } T = 2.3924, \text{ MSE} = 5.3775$$

The first confidence interval that is computed will be $\mu_A - \mu_B$, and computes out to be the following interval:

$$(-1.5534,\ 1.6174)$$

This confidence interval gives us information on the difference between the mean weight loss of Diet Group A

compared to Diet Group B. We applied the Tukey multiple, and estimated the values $\bar{Y}_A = \mu_A = 3.3000$, and $\bar{Y}_B = \mu_B = 3.2680$. The results show that the average weight loss of Diet Group A differs from that of Diet Group B in a range between -1.5534 pounds to 1.6174 pounds. We can conclude that we are 95% confident that the average weight loss for Diet Group A ranges from being less than 1.5534 pounds from the average weight loss for Diet Group B, to being more than 1.6174 pounds more than that of the average of Diet Group B. Thus, because the interval includes 0, we can further infer that there is no significant statistical difference between the average weight loss of Diet Group A, and Diet Group B.

The next confidence interval that will be computed will be that of $\mu_B - \mu_C$, and computes as the following interval:

$$(-3.5051,\ -0.4255)$$

This confidence interval gives us information on the difference between the mean weight loss of Diet Group B compared to Diet Group C. We applied the Tukey multiple, and estimated the values $\bar{Y}_B = \mu_B = 3.2680$, and $\bar{Y}_C = \mu_C = 5.2333$. The results show that the average weight loss of Diet Group B differs from that of Diet Group C in a range between -3.5051 pounds to -0.4255 pounds. We can conclude that we are 95% confident that the average weight loss for Diet Group B ranges from being less than 3.5051 pounds from the average weight loss for Diet Group C, to being less than 0.4255 pounds more than that of the average of Diet Group C. Thus, because the interval does not include, we can further infer that there is a significant statistical difference between the average weight loss of Diet Group B, and Diet Group C.

The final confidence interval computed will be $\mu_A - \mu_C$, and is computed to the following interval:

$$(-3.4897,\ -0.3769)$$

This confidence interval gives us information on the difference between the mean weight loss of Diet Group A compared to Diet Group C. We applied the Tukey multiple, and estimated the values $\bar{Y}_A = \mu_A = 3.3000$, and $\bar{Y}_C = \mu_C = 5.2333$. The results show that the average weight loss of Diet Group A differs from that of Diet Group C in a range between -3.4897 pounds to -0.3769 pounds. We can conclude that we are 95% confident that the average weight loss for Diet Group A ranges from being less than 3.4897 pounds from the average weight loss for Diet Group C, to being less than 0.3769 pounds more than that of the average of Diet Group C. Thus, because the interval does not include, we can further infer that there is a significant statistical difference between the average weight loss of Diet Group A, and Diet Group C.

Factor Effects

For an analysis of the models factor effect, we will compute a Factor Effect model given the following equation:

$$Y_{ij} = \mu + \gamma_i + \epsilon_{ij}$$

Computing this model is beneficial because it compares the difference of each group mean to that of the overall mean. For this given equation, μ, represents the overall sample mean, not separated by group. γ_i in this equation represents the effect of the diet to the group mean participating in Diet Group i. And finally, ϵ_j remains to represent the error of the jth observation in the ith group. It is important to note that the factor effect model comes with the constraint that the summation of gammas per group must sum to zero.

For calculating the gamma values, we will be subtracting the overall sample mean, denoted as (Y double bar), from each group mean. With $\mu_A = 3.3$, $\mu_B = 3.268$, $\mu_C = 5.233$, and $\mu_. = 3.9763$, this can be computed as the following formula:

$$\gamma_i = \mu_i - \mu_.$$

This is computed out to the following values:

$$\gamma_A = -0.6763, \qquad \gamma_B = -0.7083, \qquad \gamma_C = 1.2570.$$

According to the calculations above, $\gamma_A = -0.6763$, $\gamma_B = -0.7083$, and $\gamma_C = 1.2570$.

Power

Power is a value that indicates the probability of rejecting the null hypothesis given that the null hypothesis is false. This is also known as the probability of making a Type II error. For calculating power, we must calculate value ϕ. This value represents the non central parameter. The equation is given below as the following:

$$\phi = (1/\sigma_e)\sqrt{\sum n_i((\mu_i - \mu_.)^2)/a}$$

Replacing σ_e with MSE, with a value of 5.3775. With an alpha value of 0.05 With the given alpha of 0.05, and with the phi value of 2.025 , we can compute the degrees of freedom as follows:

$$\text{d.f.(num)} = a - 1, = 3 - 1, = 2, \qquad \text{d.f.(denom)} = n_T - a = 76 - 3 = 73$$

The result of the test using the power table with these given values above yields a power of 0.87 , meaning that there is an 87% probability of correctly rejecting the null hypothesis given that the null hypothesis is false.

5: Interpretation

Based on the data and analysis below, it helps people to determine if there was a statistically significant difference in weight loss between groups Diet A, Diet B, and Diet C. We can get the P-value from the ANOVA test, P-value = 0.0034, and F-Test = 6.1537. Because the P-value (0.0034) is smaller than the significance level alpha (a = 0.05), we reject the null hypothesis.

Because we reject the null hypothesis, we know that there is at least one group's mean that differs from the others. To determine which one has a difference in three groups, we need to check the confidence interval between each two of the groups. First, let's compare Group Diet A and Group B; the confidence interval between them is $\mu_A - \mu_B$ is (-1.5534, 1.6174). In this interval, we can check that 0 is in here. So between group Diet A and Diet B does not have a significant difference in average weight loss. Then, let's compare Group Diet A and Group C; the confidence interval between them is $\mu_A - \mu_C$ = (-3.4897, -0.3769). This interval suggests that participants following Diet A lost between 3.4897 and 0.3769 pounds less on average than those following Diet C. Finally, let's compare Group Diet B and Group C; the confidence interval between them is $\mu_B - \mu_C$ =(-3.5051, -0.4255). This suggests that participants following Diet B lost between 3.5051 and 0.4255 pounds less on average than those following Diet C. From this information, we can conclude that we are 95% confident there is a significant difference between average weight loss from (Group Diet A and Group Diet C) & (Group Diet B and Diet C). And we didn't find a significant difference between the average weight loss from Group Diet A and Group Diet B.

After determining these differences in average weight loss in Group Diet A, Group Diet B, and Group Diet C, we know how much each group's diet can impact the overall mean(3.9763), which we did using the gamma. Group Diet A and Group Diet B are lower than the overall mean, which is the Diet group A -0.676 and Diet group B -0.708. Group Diet C is higher than the overall mean, which is the Diet group C 1.257. In conclusion, the results of Group diet A and Group diet B are similar results; Diet C leads to significantly greater weight loss than both.

6: Conclusion

Our analysis showed that participants who followed Diet C achieved the greatest average weight loss over the six-month period. The ANOVA test indicated that at least one diet group differed significantly from the others, confirming that diet type affected weight loss.

Further confidence interval analysis showed no significant difference between Diet A and Diet B, but both had smaller mean losses than Diet C, confirming that Diet C participants lost more weight than those in the other two diets. The factor-effects model reinforced this finding, showing Diet C contributed positively to overall weight loss, while Diet A and B were slightly below the overall mean.

Diagnostic tests confirmed that model assumptions were met: no outliers were detected, residuals were approxi-

mately normal, and variances were equal across groups, which validate the reliability of our ANOVA analysis.

Finally, the power analysis showed that our test had high power, meaning a strong probability of detecting true differences among diets. This gives us confidence in rejecting the null hypothesis and supports the conclusion that Diet C is the most effective plan for losing weight, while Diets A and B produced smaller but similar effects.

Appendix:

R Code

```r
# Load data
Diet_Csv = read.csv("C:\\Users\\lderan\\Downloads\\loseit.csv")

# Load required libraries
library(ggplot2)
library(knitr)

# Histogram of Weight Loss by Diet Group
ggplot(Diet_Csv, aes(x = Loss, fill = Diet)) +
  geom_histogram(binwidth = 2, color = "black") +
  labs(
    title = "Histogram of Weight Loss by Diet Group",
    x = "Weight Loss (lbs)",
    y = "Number of Participants",
    fill = "Diet Group"
  ) +
  facet_grid(rows = vars(Diet)) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold")
  )

# Boxplot of Weight Loss by Diet Group
ggplot(data = Diet_Csv, aes(x = Diet, y = Loss, color = Diet)) +
  geom_boxplot() +
  labs(
    x = "Diet",
    y = "Weight Loss (lbs)",
    title = "Boxplot of Weight Loss by Diet Groups"
  )

# Summary statistics by group
summary_table <- Diet_Csv %>%
  group_by(Diet) %>%
  summarise(
    Sample.Size = n(),
    Mean = mean(Loss, na.rm = TRUE),
    Standard.Deviation = sd(Loss, na.rm = TRUE)
  )

# Display summary table
```

```r
knitr::kable(summary_table, digits = 4)

# Linear model and residuals
loseit.model = lm(Loss ~ Diet, data = Diet_Csv)
Diet_Csv$ei = loseit.model$residuals

# Compute MSE
nt = nrow(Diet_Csv)  # total sample size
a = length(unique(Diet_Csv$Diet))  # number of groups
SSE = sum(Diet_Csv$ei^2)  # sum of squared errors
MSE = SSE / (nt - a)  # mean squared error

# Standardized residuals
eij.star = loseit.model$residuals / sqrt(MSE)

# Outlier detection
alpha = 0.05
t.cutoff = qt(1 - alpha / (2 * nt), nt - a)
CO.eij = which(abs(eij.star) > t.cutoff)

# Studentized residuals
rij = rstandard(loseit.model)
CO.rij = which(abs(rij) > t.cutoff)

# Normality checks
qqnorm(loseit.model$residuals)
qqline(loseit.model$residuals)

# Shapiro-Wilk test
ei = loseit.model$residuals
loseit.SWtest = shapiro.test(ei)
loseit.SWtest

# Residual plots
plot(
  loseit.model$fitted.values, loseit.model$residuals,
  main = "Errors vs. Group Means",
  xlab = "Group Means",
  abline(h = 0, col = "purple")
)

# Boxplot of residuals
boxplot(ei ~ Diet, data = Diet_Csv)

# Equality of variance test
library(car)
the.BFtest = leveneTest(ei ~ Diet, data = Diet_Csv, center = median)

# Extract p-value
p.val = the.BFtest[[3]][1]
p.val
```