STA 106 - Project 1

Matthew Holcomb, Runyu Yang, Adam Hetherwick

Instructor: Maxime Guiffo Pouokam

2/14/2023

The Impact of Diet (A, B, C) on Losing Weight

Topic One: The Impact of Diet (A, B, C) on Losing Weight

I. Introduction

In this modern society, an increasing number of people are concerned about weight, and people are starting to look at which diets help them lose weight the most. The aim of the study is to determine which diet is most effective for weight loss by comparing the weight loss results of three different diets (diet A, B and C). Generally, the loseit dataset is an observational study, and response variable is weight loss, in pounds, which is the difference between their weight at the beginning of the program, and their weight after 6 months, while the explanatory variable is the diet used. Using data from 76 participants, the study will compare the weight loss between the diets using the ANOVA group means model, provide pairwise confidence intervals for differences in means, and investigate the contribution of each group to the overall mean. The initial question that we will emphasize is if there is statistical evidence to show that the true population average weight loss is the same for all 3 types of diet (A, B and C), using a 5% significance level ($\alpha = 0.05$). After considering this question, we can generally conclude which diet, if any, will have an impact on weight loss. To achieve this, we will perform the F-test as part of the analysis of variance to test the null hypothesis that the mean weight loss for all three diets are equal. If the ANOVA test results in a p-value less than 0.05, we will reject the null hypothesis and conclude that at least one sample mean is different from others. However, we would fail to reject the null hypothesis when p-value is more than 0.05.

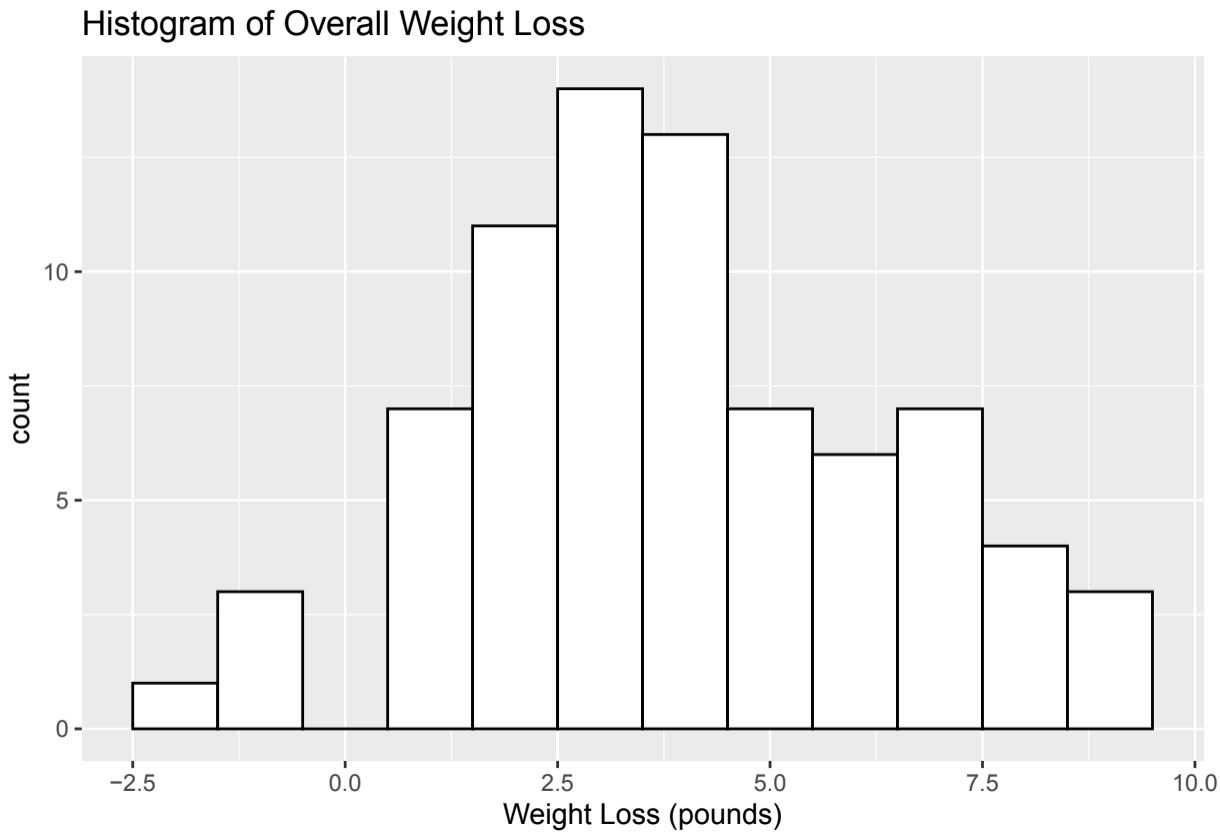
In this project, we have another goal of providing paired confidence intervals for mean differences, which will allow us to estimate the possible range of the true mean differences between diets based on sample data. In addition, we will compare weight loss outcomes between

diets and determine the contribution of each group's mean to overall mean using a “factor effect” model to investigate the contribution of each group mean (γi) to the overall mean. To accomplish this, we will construct the confidence interval at 95% that compare one group mean to another, including $\mu_A - \mu_B$, $\mu_A - \mu_C$, and $\mu_B - \mu_C$. The intervals constructed for these differences represent the possible range of true differences between the two sets of means based on the sample data, with 95% confidence. The sample mean, standard deviation, and sample size for both groups were used to calculate the interval. If the confidence interval for mean difference between the two diets does not contain zero, we can conclude with 95% confidence that the true mean difference between the two groups is significantly different. This information will give us a higher level of certainty about the results of our analysis and allow us to make more informed conclusions about the efficacy of each diet, such as which one produces the best results (which in this case means the most weight lost over the six month period). The results of this study will provide valuable insights into the effectiveness of different diets and help promote a healthier and more sustainable approach to weight loss.

II. Summary of Data

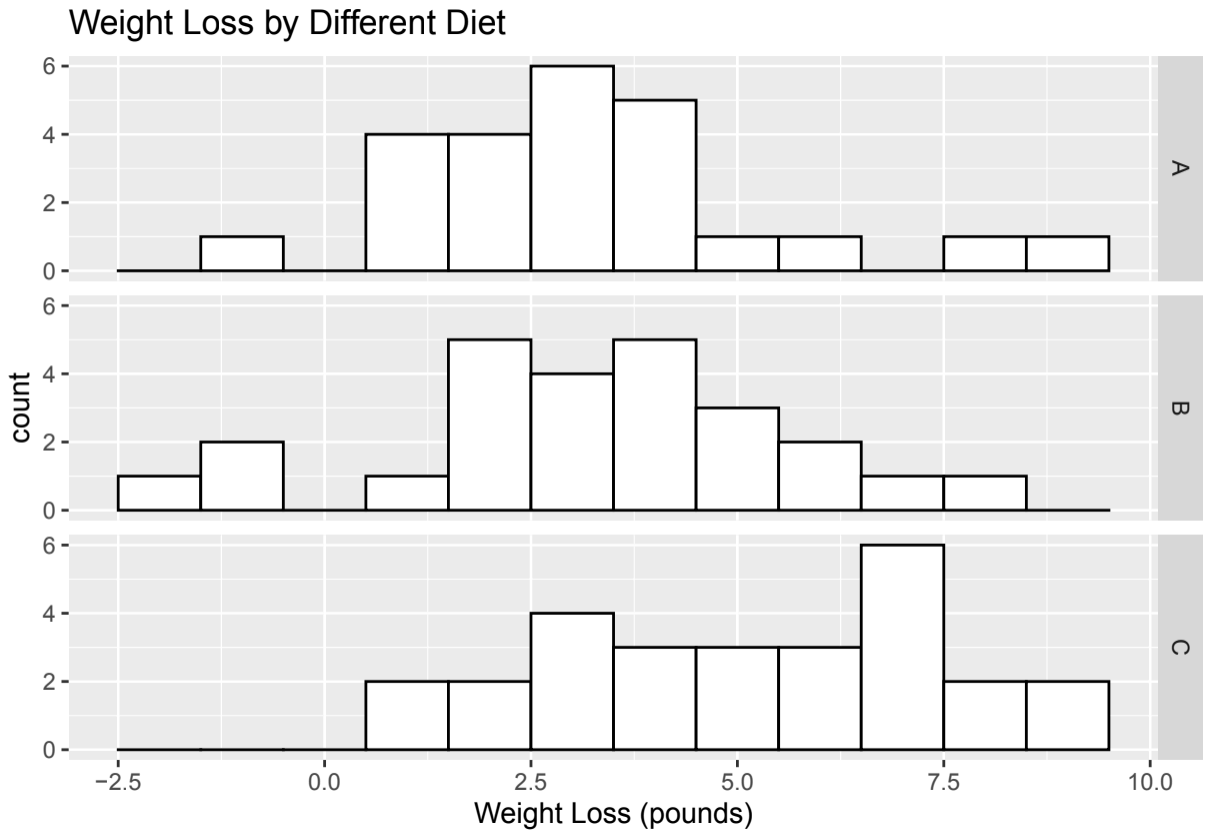
The graphs and tables listed below assist us in visualizing the weight loss data and help us analyze the change in weight after 6 months of a particular diet (A, B, or C). The graphs do not offer much quantitative or numeric representation to solidify concrete statistical conclusions though they do represent the weight loss visually. With our graphs, we can understand the mean distribution by diet, and make predictions which the regression analysis will further support or contradict. We can also observe skew and identify outliers in the dataset, for example, people who either gained a bit of weight from the diet (unusual outcome) or people who lost nearly

significantly more weight than most others (over twice as much in some cases!). We can use this information to compare the average weight loss over 6 months, with the hope of identifying which of the three diets (if any) produce the best or worst results.

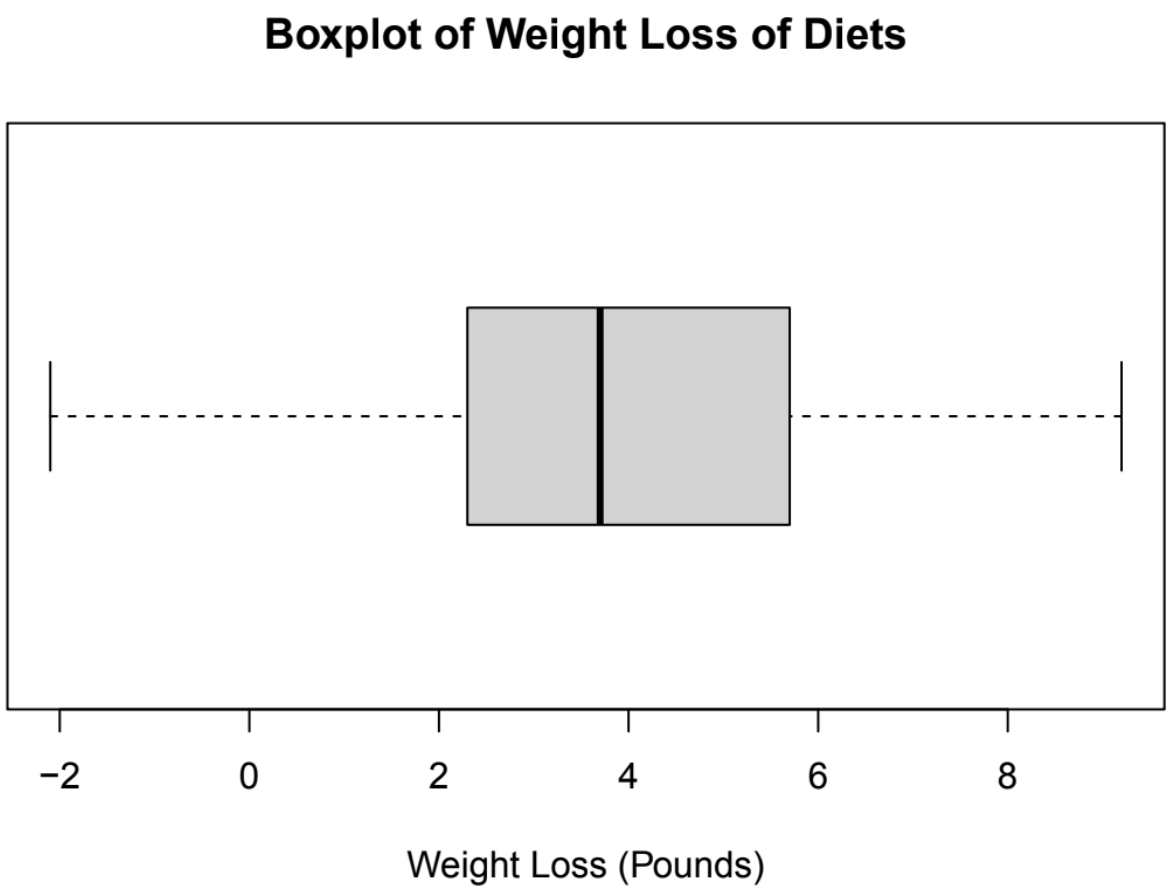


This histogram displays the weight loss distribution of the 76 people who took diets A, B, or C. The height of each bar represents the frequency or count of that bin. The higher the bar, the more common that bin is. Bins below 0 represent people who gained weight over the 6 months while following the diet. The data appears to be normally distributed with mean 3.9763lbs and standard deviation 2.4732lbs. While the data appears to be normally distributed, it does seem slightly right-skewed, which suggests that those who lost more weight than average have a more

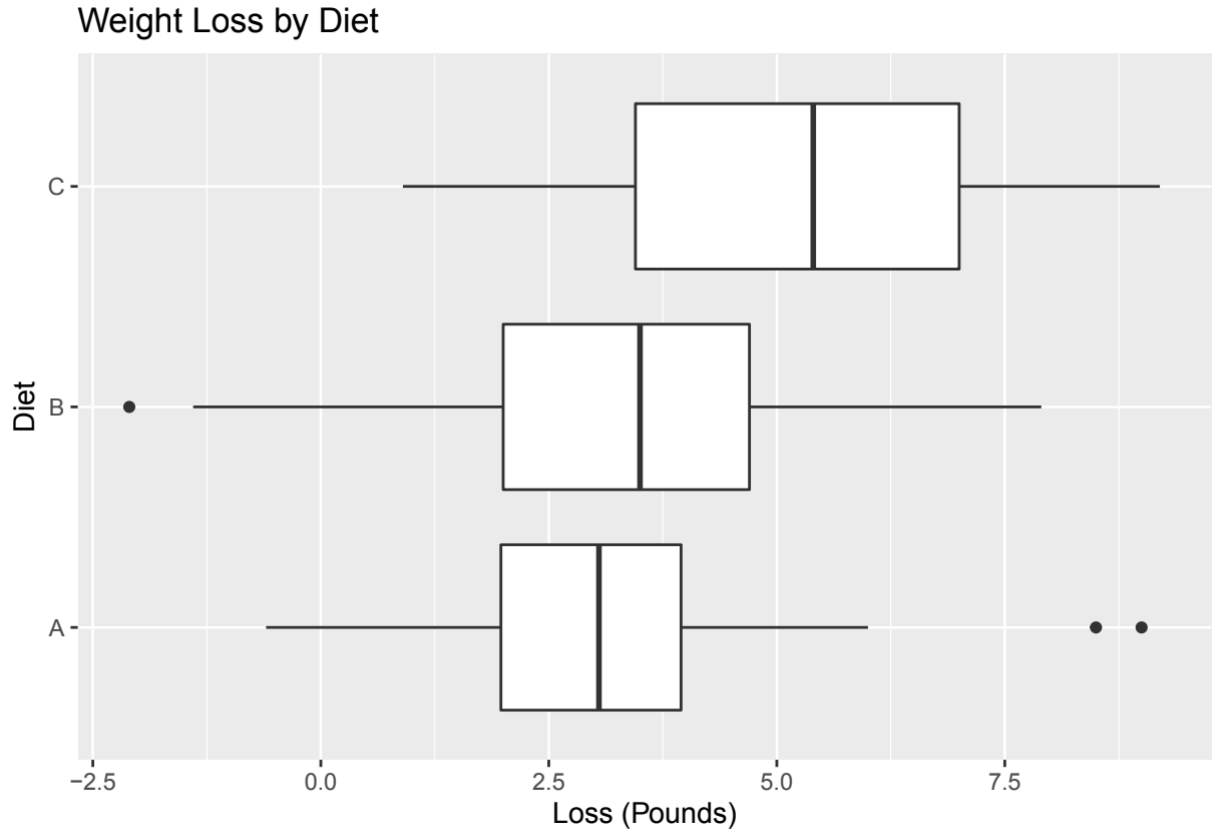
condensed distribution as opposed to those who lost less or gained weight, where those individual's (less weight loss compared to the mean) results are more spread out.



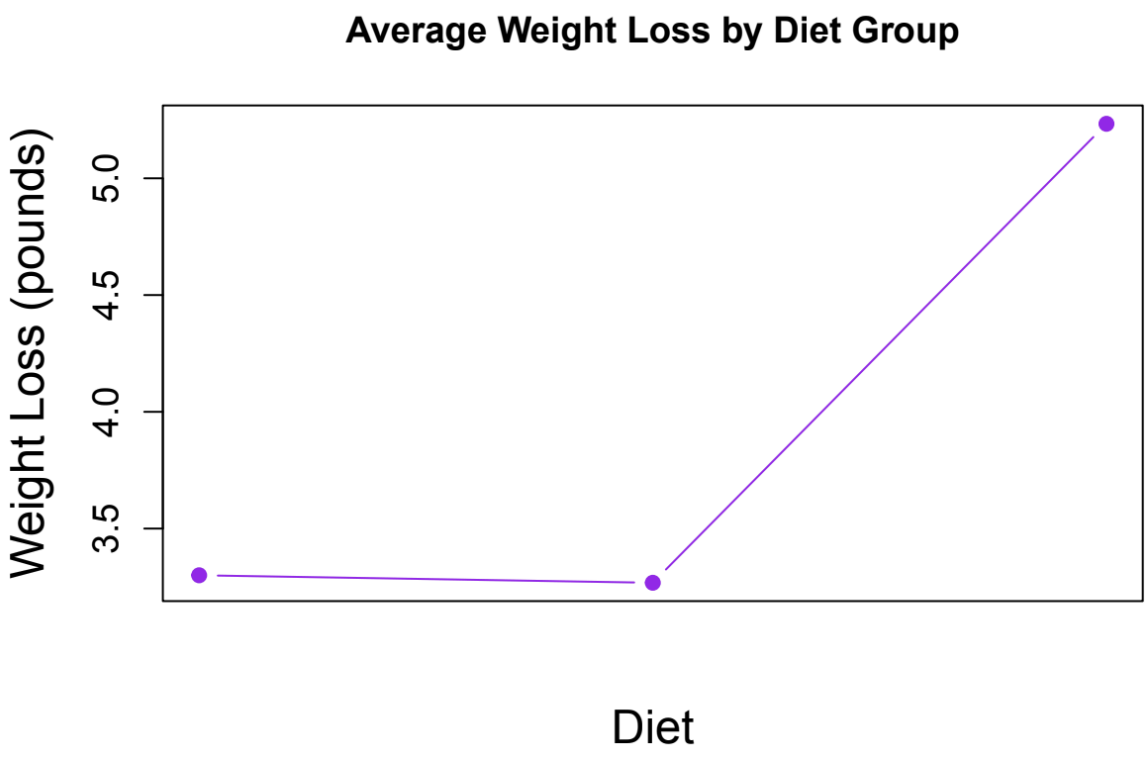
These 3 graphs represent the weight difference after 6 months of using a certain diet. Each histogram appears to be normally distributed with means 3.3lbs for diet A, 3.268lbs for diet B, and 5.23lbs for diet C. We can see from the graph that each distribution around the mean lbs lost appears to be similar, with standard deviations 2.2401lbs for diet A, 2.4645lbs for diet B, and 2.2477lbs for diet C. We notice that in diets A and B, there are people that gained weight from the diet, as seen by the frequency bar below 0. The vast majority of each diet group lost weight.



This graph shows the distribution of weight loss for all the diets. As we can see, there are no outliers in the dataset. The black bar in the middle of the center around 3.9 represents the mean average weight loss in lbs. The bars near 2 and 6 represent the upper and lower quartiles, which signifies where most of the data lies. This means the most common weight loss in lbs is around 2.1lbs and 4.9lbs. There is relatively high variation within the dataset, as seen by the minimum weight lost at around -2 and the maximum weight lost around 10.



These boxplots show the distribution of each weight loss by diet (A, B, or C). As we can see from the top boxplot, diet C had the highest weight loss and had no outliers. From the second boxplot signifying diet B we see the largest spread, ranging from people who lost up to 7.5lbs on the diet, to people who gained weight on the diet. Group B's mean weight loss was still lower than group A, which had smaller variability but still had lower mean weight loss. Diets A and B had outliers as signified by the dots outside of the quartile ranges. Diet B had an outlier who gained weight and diet A had outliers that lost so much weight they were considered outliers.



This graph represents the mean weight loss by diet (Group A being the far left point, Group B in the middle, and Group C on the right). As we can see, diets A and B had means just below 3.5 while diet C had a mean above 5. This graph gives us a good representation of what we may expect to get from our hypothesis test later on that the group means are not equal to each other. Since diet C is so much different than diet A or B, we may expect to reject the null and conclude that one mean is significantly different from the others.

SUMMARY DATA	GROUP A	GROUP B	GROUP C	OVERALL
MIN	-0.6000	-2.1000	0.9000	-2.1000
Q1	1.9500	2.0000	3.4500	2.3000
MEDIAN	3.0500	3.5000	5.4000	3.7000
Q3	4.0000	4.7000	7.0000	5.7000
MAX	9.0000	7.9000	9.2000	9.2000
MEAN	3.3000	3.2680	5.2333	3.9763
STD DEVIATION	2.2401	2.4645	2.2477	2.4732
SIZE	24	25	27	76

This table provides us with numeric values that we can use to calculate statistics on the dataset. We can see the minimum average values for each diet, with B being the lowest. We can see the interquartile range from Q1 to Q3 for each diet, with the median included as well. One thing we noticed from this graph was that the overall median is much lower than the diet C median and it is closer to that of diet B and A, thus signifying diet C may be significantly different from that of diet B or A. We can also observe that the standard deviations of each group are roughly the same, which indicates that the errors of our sample are normally distributed. That is, the amount that each group's samples vary from the group mean appears to be approximately equal. This is important, as the assumption of equal variances must be met in order to conduct our analysis using ANOVA.

III. Analysis

III.1 Single Factor ANOVA Group Means Model

We will use the single factor ANOVA group means model for our analysis:

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

Using single factor ANOVA takes into account the following assumptions:

(1) Y_{ij} are randomly sampled

(2) Errors are normally distributed $\epsilon_{ij} \sim N(0, \sigma_{\epsilon}^2)$

(3) The *i* groups are independent

We want to determine whether the different diets (A, B, and C) have an impact on the weight loss of participants during the six month program. To do this, we will compare the three groups' mean weight loss, testing to see if a statistically significant difference exists between them. Using single factor ANOVA, our null hypothesis will represent the statement "there is no difference in weight loss between each group," while our alternative hypothesis will state "at least one of our groups will have a difference in weight loss compared to the others." These hypotheses are mathematically defined as:

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

Alternative hypothesis: H_A: $At\ least\ one\ \mu_A,\ \mu_B,\ \mu_C \neq$

Note: Going forwards, group A will interchangeably be referred to as group 1, B as group 2, and C as group 3. This is important notation-wise, as we will refer to "the i^{th} group" frequently, which in turn means if $i = 1$, we are talking about group A, etc.

Since we do not know the *true* average weight loss for diet groups A, B, and C, we will use our sample data to estimate them, where average weight loss, μ_i, is estimated by the sample mean for each group, $\overline{Y}_{i\bullet}$. Similarly, we will estimate the true population variance, ϵ_{ij}, using its sample estimator *MSE*.

III.2 F-test for ANOVA

As mentioned earlier, group C seems to have a larger average weight loss compared to groups A and B, as seen in this simplified table:

SAMPLE INFO	GROUP A	GROUP B	GROUP C	OVERALL
MEAN WEIGHT LOSS (POUNDS)	3.3000	3.2680	5.2333	3.9763
STD. DEVIATION (POUNDS)	2.2401	2.4645	2.2477	2.4732
SAMPLE SIZE	24	25	27	76

We will conduct a F-test in order to mathematically determine whether there is in fact some statistically significant difference between the group means for each diet group. Our test statistic is given by $F_S = \frac{MSA}{MSE}$, where *MSA* is the mean of the sum of squared errors for all group means (A, B, C) compared to the overall mean (regardless of group), and *MSE* is the mean of the sum of squared errors of each observation compared to their group's sample mean. We calculate the test statistic and compare it to the F distribution, operating under the null assumption of there being no difference between means. The larger the test statistic is, the larger the difference between the *MSA* and *MSE*, which in turn suggests higher variation between the variance of individual groups compared to the overall variance: if there was no difference in means, we would expect data from each group to deviate from the group mean similarly as it does to the overall mean, since they would all be equal.

Using R to produce the ANOVA table, we calculated *MSE* = 5.3775, *MSA* = 33.0915, and our test statistic F_s = 6.1537:

ANOVA	DF	SUM SQ	MEAN SQ	F VALUE	PR(>F)
DIET	2	66.18	33.091	6.1537	0.00339
RESIDUALS	73	392.55	5.377		

As mentioned, our test statistic is then compared to the F distribution with degrees of freedom for the numerator a - 1 (a being the number of groups; in this case a = 3), and degrees of freedom for the denominator n_T - a (n_T being the overall number of subjects sampled; n_T = 76). These values can be seen as the DF for Diet (numerator) and Residuals (denominator) in the ANOVA table above.

This produces the p-value, $Pr\{F > F_s\}$, for our test statistic, 0.0034. This value represents that the probability of, assuming the null hypothesis is true, sampling the same data or more extreme is 0.0034. Using $\alpha = 0.05$, we find that our p-value < α, leading us to reject the null hypothesis and conclude that there is a statistically significant difference between at least one of the group means. This means that at least one of the diet programs leads to a different amount of weight loss compared to the others. However, we are unable to determine which specific diet(s) differ in results from just our F-test.

III.3 Pairwise Confidence Intervals for Differences in Means

In order to identify which group means differ from the others, we can construct pairwise confidence intervals at a 95% confidence level that compare specific groups to one another. To compare two independent means, we use the equation:

$$(\overline{Y}_{i\bullet} - \overline{Y}_{i'\bullet}) \pm t_{1-\frac{\alpha}{2}, n_T - a} \times \sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_{i'}})}$$

As mentioned, we want to compare each of our means to one another, so we would plug in the comparison of group A to group B as $i = 1$ and $i' = 2$, etc.

Note: $\alpha = 0.05,\ n_T = 76,\ a = 3,\ MSE = 5.3775,\ n_i = \#\ subjects\ in\ group\ i,$

$\overline{Y}_{i\bullet}$ represents mean weight loss for i^{th} group ($i = 1$: diet A, etc.)

We can use R to calculate each the confidence intervals for $\mu_A - \mu_B, \mu_A - \mu_C, \mu_B - \mu_C$, where our interval represents a range of values that we are 95% confident the true difference between the compared means falls within:

PAIRWISE 95% CONFIDENCE INTERVALS	MINIMUM BOUND	ESTIMATE	MAXIMUM BOUND
$\mu_A - \mu_B$	-1.2887	0.0320	1.3527
$\mu_A - \mu_C$	-3.2299	-1.9333	-0.6368
$\mu_B - \mu_C$	-3.2481	-1.9653	-0.6826

From these intervals, we can draw further conclusions than what we determined using our F-test. Since the interval comparing groups A and B contains 0, we are unable to identify any difference between the true group means of diets A and B. However, when comparing groups A and B to C, we find that both intervals do not include zero, which allows us to conclude that there is a difference in weight loss between groups A and C as well as between groups B and C. Specifically, we are 95% confident that the true average weight loss for those following diet C is between 0.6368 and 3.2299 pounds more than those following diet A, while true weight loss for those following diet C is between 0.6826 and 3.2481 pounds more than for those following diet B.

III.4 Power Calculations

The power of a test is the probability that we will correctly reject the null hypothesis. In order to calculate power, we assume that the alternative hypothesis is true, and calculate ϕ using the following equation:

$$\phi = \frac{1}{\sigma_\epsilon}\sqrt{\frac{\sum_{i=1}^{n} n_i(\mu_i - \mu_\bullet)^2}{a}}$$

Note: ϕ is the non-central-F parameter, which measures how different the test statistic F_s is when considering the alternative hypothesis to be true as opposed to when we consider the null

hypothesis true. Since we do not know the actual values of the group means and overall standard deviation, we must estimate them:

σ_ϵ is estimated using $\sqrt{MSE}$

μ_i and $\mu_\bullet$ is estimated using $\overline{Y}_{i\bullet}$ and $\overline{Y}_{\bullet\bullet}$

We are evaluating at $\alpha = 0.05$, with degrees of freedom: d.f.{num} = 2, d.f.{denom} = 73. Plugging these values into the equation above allow us to calculate ϕ, which in turn is used to calculate the power of our test using R. Doing so gives power = 0.8778, meaning that the estimated probability that we correctly reject the null hypothesis, that the average weight loss for each of the three diet groups is the same, while in reality there *is* a difference in average weight loss between groups, is 0.8778.

III.5 Contribution of Group Means to the Overall Mean Using "Factor Effects"

Another way to model how *Y* changes with one categorical variable is using the alternative ANOVA model:

$$Y_{ij} = \mu_\bullet + \gamma_i + \epsilon_{ij}$$

We no longer consider the i^{th} group mean μ_i, instead using the "effect" of the i^{th} group γ_i, which represents how the overall mean is affected by a subject being in group *i*. The assumption for this type of ANOVA are the same as single factor, except we include the constraint:

$$\sum_{i=1}^{a} \gamma_i = 0$$

Once again, we do not know the group "effects" γ_i, so we estimate them using $\hat{\gamma}_i$, which is calculated using:

$$\hat{\gamma}_i = \overline{Y}_{i\bullet} - \overline{\overline{Y}}, \text{ where } \overline{\overline{Y}} = \frac{1}{a}\sum_{i=1}^{a} \overline{Y}_{i\bullet}$$

We can calculate the estimated "effects" using R, where we directly compare the group means to the overall mean, and find:

$$\hat{\gamma}_1 = -0.6338, \hat{\gamma}_2 = -0.6658, \hat{\gamma}_3 = 1.300$$

This gives us the estimated impact on the overall average weight loss by diet group, where diet A results in an average of 0.6338 fewer pounds lost compared to the overall mean, diet B results in an average of 0.6658 fewer, and diet C results in 1.300 more pounds lost, all over the six month period. This is comparable to our results from the pairwise confidence intervals constructed earlier, where we were unable to find a difference between diets A and B, but found that there was a statistically significant increase in weight loss using diets A or B compared to diet C. We can see that the effect on the overall mean when participants used diets A or B were very similar, while diet C was quite different from both.

IV. Interpretation

IV.1 F-test results

As we can see from the results, the F-test result is that there is a statistically significant difference between the group means of the diets (A, B, C), and that at least one of the diets leads to a different amount of weight loss compared to the others. The test statistic Fs (6.1537) is calculated as the ratio of the mean of the sum of squared errors for all group means (MSA) to the mean of the sum of squared errors of each observation (MSE). This test statistic is then compared to the F distribution with appropriate degrees of freedom to produce a p-value (0.0034), which represents the probability of observing the same or more extreme data if the null

hypothesis (no difference between means) is true. The p-value is found to be less than the significance level (0.05), leading us to reject the null hypothesis and conclude that there is a statistically significant difference between at least one of the diet groups. However, it is important to note that the F-test does not reveal which specific diets differ in their results.

IV.2 Pairwise Confidence Interval Results

The pairwise confidence intervals further elaborate on which group means are different. The intervals for A - B contain zero, meaning that there is no difference between the true group means of diets A and B. However, the intervals for A - C and B - C do not contain zero, indicating that there is a difference in weight loss between group A and group C, as well as between group B and group C. The confidence intervals show that we can be 95% confident that the true average weight loss for those following diet C is between 0.6368 and 3.2299 pounds more than those following diet A, while true weight loss for those following diet C is between 0.6826 and 3.2481 pounds more than for those following diet B. Based on the results, it can be seen that diet C leads to more pounds lost on average compared to the overall average, while diet A or B lead to fewer pounds lost compared to the overall average. The results of this analysis were consistent with those obtained from pairing confidence intervals, with significant differences found between diets A or B and C, but not between diets A and B. The estimated impact of each diet group on the overall mean showed that diet A and diet B had similar effects on weight loss, while diet C had significantly different effects.

IV.3 Contribution of Group Means ANOVA Model

The results from using the contribution of the group means ANOVA model aligned with the results from the pairwise confidence intervals as discussed above. We found that there is a positive effect of diet C on the overall mean weight loss, where those following diet C lost 1.300

pounds more than average for the whole sample, while those following diets A and B lost roughly 0.67 pounds less than average. This supports the idea that diets A and B have similar, less effective results compared to diet C, where once again it is clear that diet C is the best of the three for weight loss over a six month period.

V. Conclusion

Our original analysis of the group means suggested that diet C led to more weight loss over a six month period when compared to diets A and B. We constructed an ANOVA model for our data, and conducted a F test which determined that there was in fact a statistically significant difference in at least one of our group means; that at least one of the diets had a different impact than the others. In order to identify which diets performed better or worse compared to one another, we constructed pairwise confidence intervals that compared each diet directly to the others. We found that the weight loss was consistent between participants following diets A and B, but that those following diet C lost more weight over the six month period than those following diets A or B. We then compared each diet's average weight loss to the overall mean weight loss, regardless of group, using the "Factor Effects" ANOVA model, and found comparable results: diets A and B produced similar, below-average (when compared to the average weight loss of a participant using one of the three diets) weight loss, while diet C produced above-average weight loss. These analyses allow us to conclude that, between the three, diet C produces the most weight loss, while diets A and B are statistically equivalent in the amount of weight loss produced.