ANOVA

The a previous lab we introduced the two-group independent t-test as a method for comparing the means of two groups. In some settings, it is useful to compare the means across more than two groups. The methodology behind a two-group independent t-test can be generalized to a procedure called **analysis of variance (ANOVA)**. Assessing whether the means across several groups are equal by conducting a single hypothesis test rather than multiple two-sample tests is important for controlling the overall Type I error rate.

The material in this lab corresponds to Section 7.5 of *OpenIntro Statistics*.

FAMuSS: comparing change in non-dominant arm strength by *ACTN3* genotype

Is change in non-dominant arm strength after resistance training associated with genotype?

In the Functional polymorphisms Associated with Human Muscle Size and Strength study (FAMuSS), researchers examined the relationship between muscle strength and genotype at a particular location on the *ACTN3* gene. The `famuss` dataset loaded below contains a subset of data from the study.

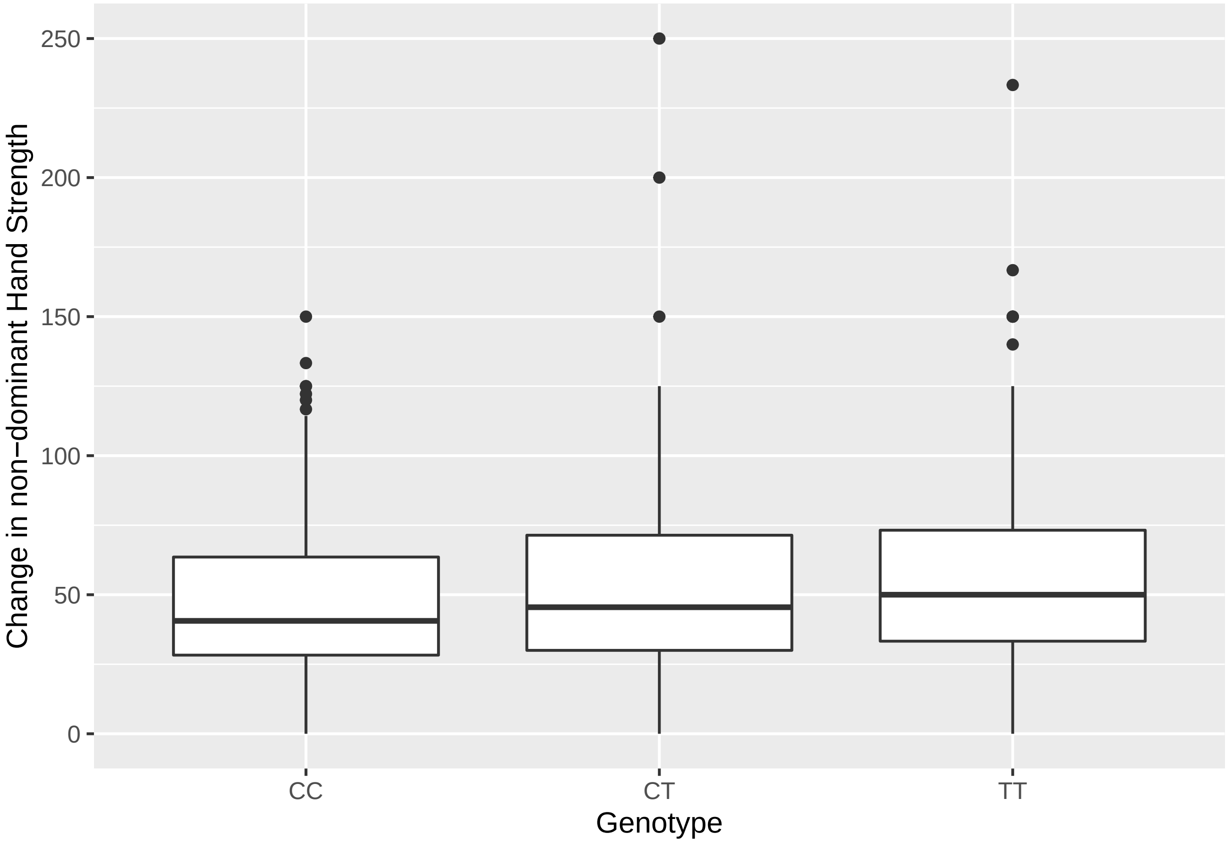
The percent change in non-dominant arm strength, comparing strength after resistance training to before training, is stored as `ndrm.ch`. There are three possible genotypes (CC, CT, TT) at the *r577x* position on the *ACTN3* gene; genotype is stored as `actn3.r577x`.

1. Load the data. Create a plot that shows the association between change in non-dominant arm strength and *ACTN3* genotype. Describe what you see.

WJ Response:

```
#load the data
load(url('https://github.com/jbryer/DATA606/blob/master/data/famuss.rda?raw=true'))
#create plot
plt_data <- famuss %>%
  filter(!is.na(actn3_r577x))

ggplot(data = plt_data, aes(x = actn3_r577x, y = NDRM.CH)) +
  geom_boxplot() +
    labs(x = 'Genotype',
         y = 'Change in non-dominant Hand Strength'
         )
```



Visual analysis of the boxplot above reveals that there are slight differences in the medians and spread for each genotype. This indicates that it might be an influencing factor in the percent change of non-dominant hand strength.

---

2. Assess whether the assumptions for conducting an ANOVA are reasonably satisfied: 1) observations are independent within and across groups, 2) the data within each group are nearly normal, and 3) the variability across the groups is about equal.

WJ Response:

Observations are independent within and across groups…

A quick Google search reveals that the data in this survey is random and we are assuming the sample we are using is also. Additionally, a quick check shows that for each case there is only a single genotype of the ACTN3 gene:

```
unique(plt_data$actn3_r577x)
```

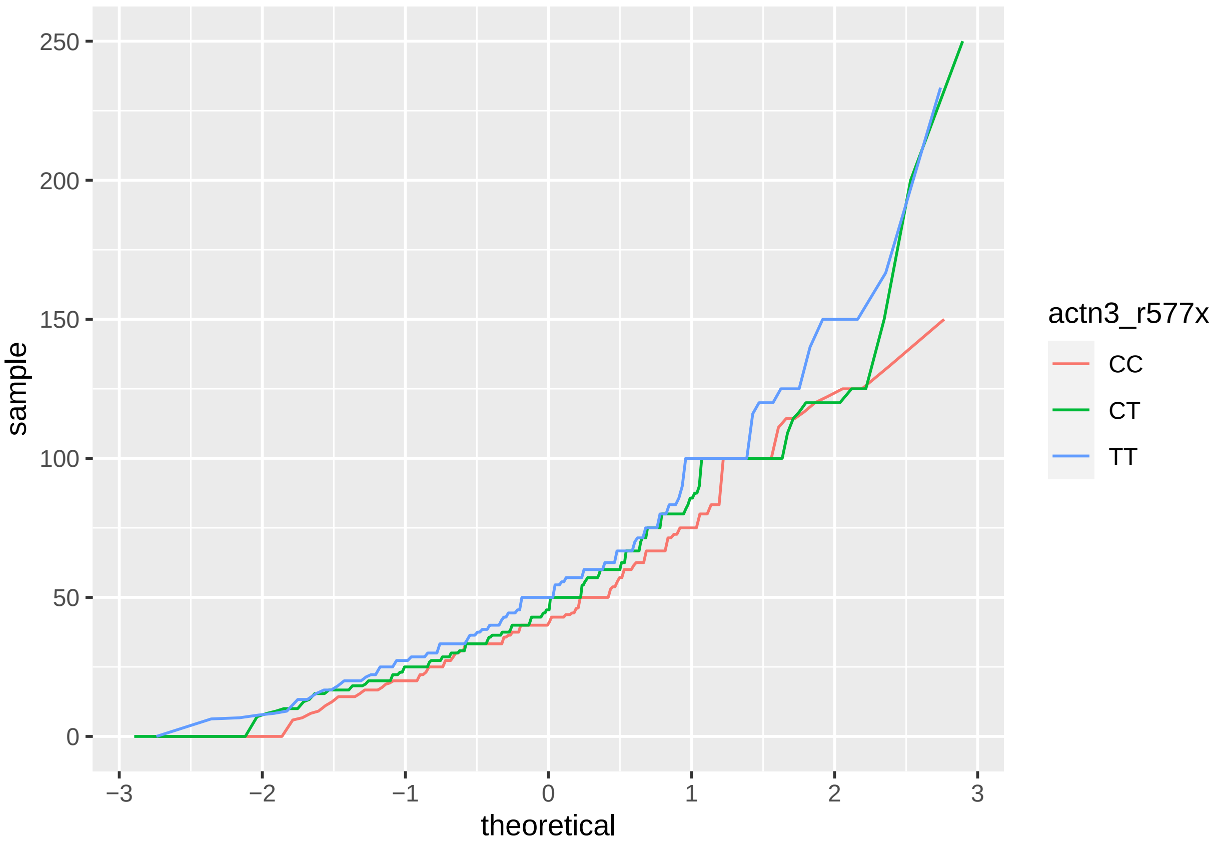
```
## [1] "CC" "CT" "TT"
```

This confirms that for our data there is independence both between and within groups.

The data within each group are nearly normal…

The following code chunk generates Q-Q plots for the cases within each group (each genotype):

```
ggplot(data = plt_data,
       aes(sample = NDRM.CH, group=actn3_r577x, colour=actn3_r577x)) +
  geom_line(stat = "qq", )
```



While there is some deviation from a line-like shape towards the ends of the plot, the middle values seem to approximate a simple sloped line. Given this, we can assume that the data within each group is approximately normal.

The variability across the groups is about equal

The below cell calculates the variance for each of the groups:

```
plt_data %>%
  group_by(actn3_r577x) %>%
    summarize(group_variance = var(NDRM.CH, na.rm = TRUE))
```

```
## # A tibble: 3 x 2
##   actn3_r577x group_variance
##   <chr>                <dbl>
## 1 CC                    905.
## 2 CT                   1103.
## 3 TT                   1259.
```

These variances are all approximately equal, with the ratio of the largest variance to the smallest being about 1.4.

Next, let's conduct a hypothesis test to address the question of interest. Let $\alpha = 0.05$.

a) Let the parameters μ_{CC}, μ_{CT}, and μ_{TT} represent the population mean change in non-dominant arm strength for individuals of the corresponding genotype. State the null and alternative hypotheses.

WJ Response:

The null and alternative hypotheses in this case are:

$$\begin{aligned} H_0 :& \mu_{CC} = \mu_{CT} = \mu_{TT} \\ H_1 :& \exists i, j \mid \mu_i \neq \mu_j \end{aligned}$$

where i or j represent any of CC, CT, or TT.

---

b) Use `summary(aov())` to compute the F-statistic and p-value. Interpret the p-value.

WJ Response:

The F-statistic and p-value are calculated in the cell below:

```r
aov_mod <- aov(NDRM.CH ~ actn3_r577x, data = famuss)
summary(aov_mod)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## actn3_r577x   2   8161    4081   3.753  0.024 *
## Residuals   600 652359    1087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 794 observations deleted due to missingness
```

Given that the p value in this case is less than 0.05, we can reject the null hypothesis and conclude that at least one pair of genotypes has statistically different mean values of the percent change in non-dominant hand strength.

Next, we will complete the analysis using pairwise comparisons.

i. What is the appropriate significance level $\alpha^\star$ for the individual comparisons, as per the Bonferroni correction?

$$\alpha^\star = \alpha/K, \text{where } K = \frac{k(k-1)}{2}\text{for k groups}$$

WJ Response:

The following code cell translates the above formula into R code for our specific use case:

```r
num_groups <- length(unique(plt_data$actn3_r577x))
alpha_i <- 0.05

K <- num_groups * (num_groups - 1) /2
alpha_f <- alpha_i / K
alpha_f
```

```
## [1] 0.01666667
```

Thus, the appropriate α level in this case is 0.16667.

---

ii. Use `pairwise.t.test()` to conduct the pairwise two-sample t-tests.

WJ Response:

```
clean_df <- famuss %>%
  filter(!(is.na(ND23.CH) | is.na(actn3_r577x)))

pairwise.t.test(clean_df$NDRM.CH, clean_df$actn3_r577x)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  clean_df$NDRM.CH and clean_df$actn3_r577x
##
##    CC   CT
## CT 0.19 -
## TT 0.03 0.24
##
## P value adjustment method: holm
```

```
pairwise.t.test(clean_df$NDRM.CH, clean_df$actn3_r577x, p.adj = 'bonf')
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  clean_df$NDRM.CH and clean_df$actn3_r577x
##
##    CC   CT
## CT 0.28 -
## TT 0.03 0.73
##
## P value adjustment method: bonferroni
```

iii. Summarize the results.

WJ Response:

Using our original α value of 0.05, we can see that in both instances the only statistically significant difference between means occurred when comparing the CC and TT genotypes. However, since we are now looking at the p-values within groups, we now have to use our adjusted α^* to test for significance. This has the effect of "spreading" our significance over each group. Using α^* none of the paired t tests resulted in a statistically significant result. We do see that when comparing the results of the t tests that used the Bonferroni correction to those that did not, two of the p values increased. This makes sense, seeing as the point of the Bonferroni is to reduce the chance that we make a Type I error. In this case, it worked!

NHANES: comparing BMI by educational level

Is body mass index (BMI) associated with educational attainment?

This section uses data from the National Health and Nutrition Examination Survey (NHANES), a survey conducted annually by the US Centers for Disease Control (CDC). The dataset `nhanes.samp.adult.500` contains data for 500 participants ages 21 years or older that were randomly sampled from the complete NHANES dataset that contains 10,000 observations.

The variable `BMI` contains BMI information for the study participants. The variable `Education` records the highest level of education obtained: 8th grade, 9th - 11th grade, high school, some college, or college degree.

4. Load the data. Create a plot that shows the association between BMI and educational level. Describe what you see.

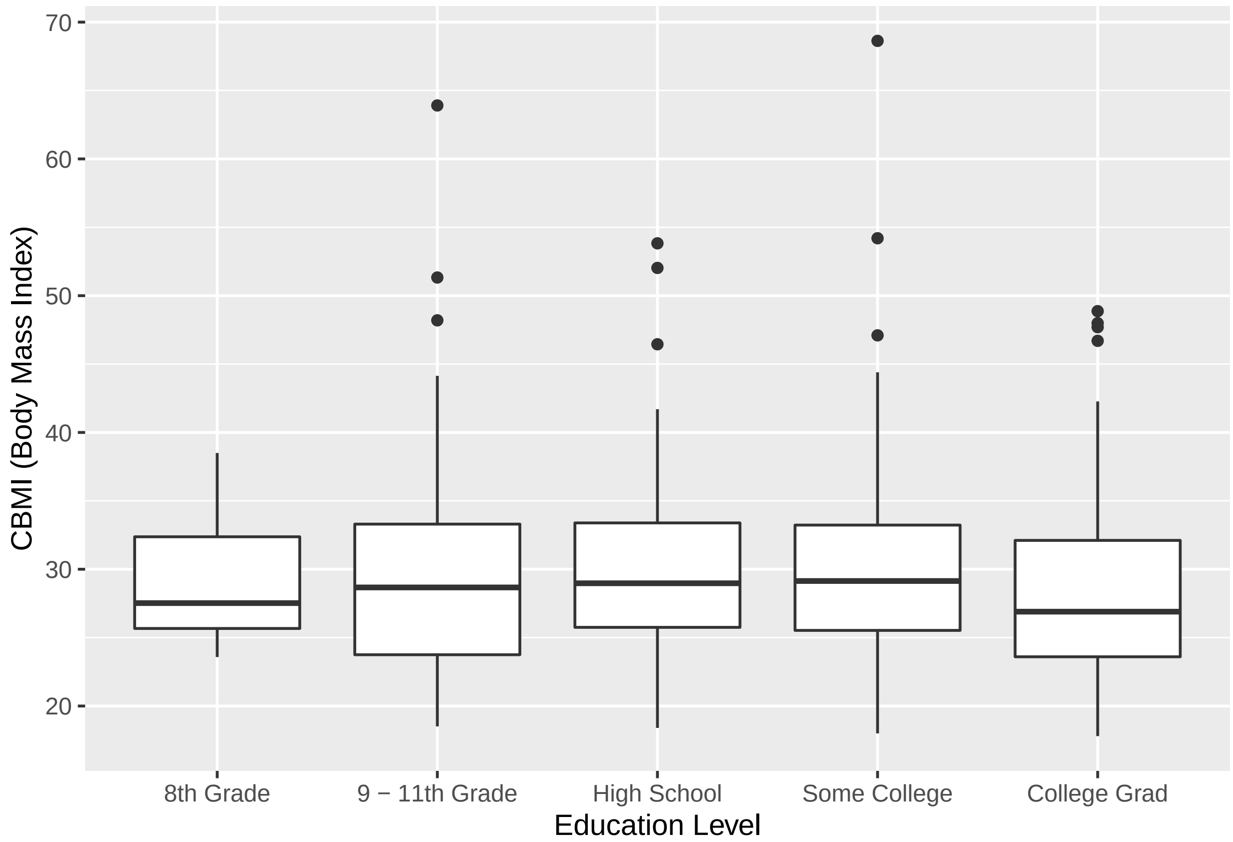
WJ Response:

```
#load the data
load(url('https://github.com/jbryer/DATA606/blob/master/data/nhanes_samp_adult_500.rda?raw=true'))

nhanes <- nhanes.samp.adult.500

#create a plot
clean_data <- nhanes %>%
  filter(!(is.na(BMI) | is.na(Education)))

ggplot(data = clean_data, aes(x = Education, y = BMI)) +
  geom_boxplot() +
    labs(x = 'Education Level',
         y = 'CBMI (Body Mass Index)'
         )
```



The histogram above reveals that there might be some groups that have different spreads/medians of BMI values, and thus that there might be a relationship between education level and BMI. However, further testing must be done to prove if any of these groups have any statistically significant differences.

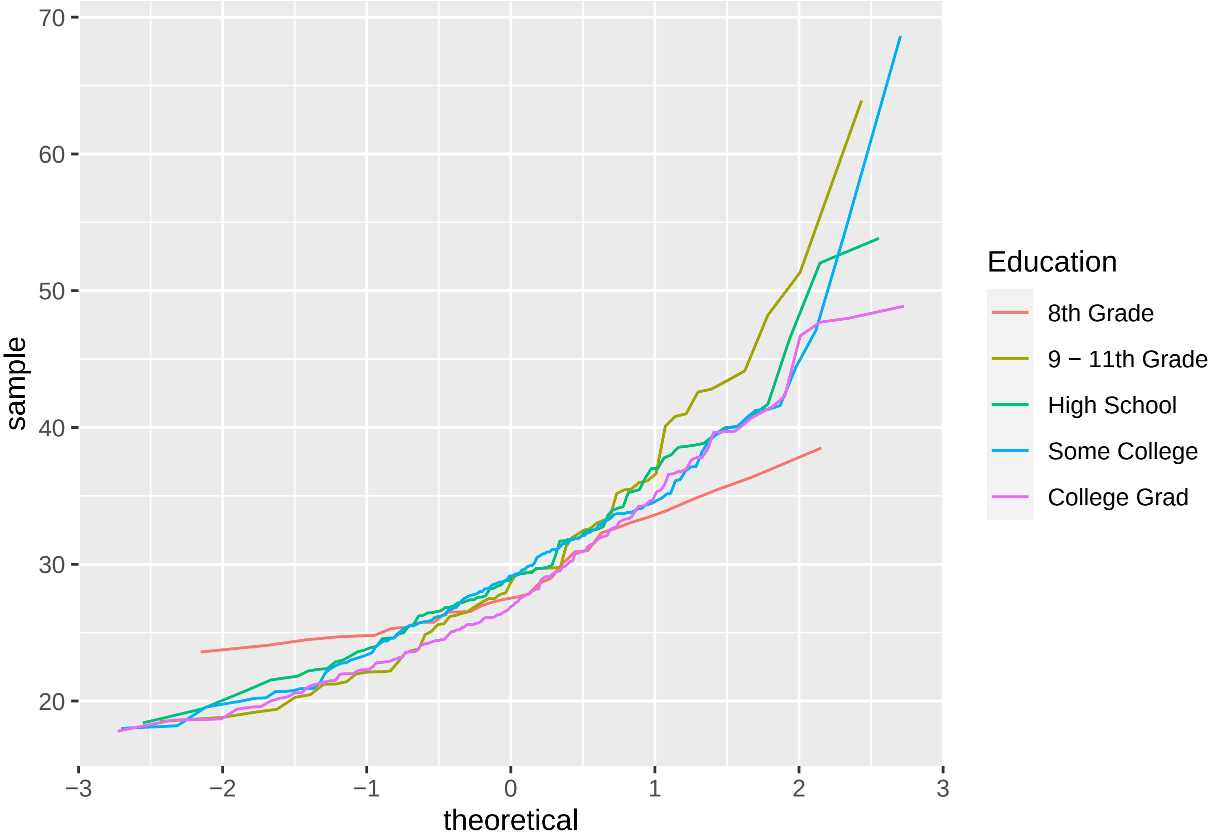
---

5. Examine the normality and equal variance assumptions across the groups. Explain why it is advisable to restrict the analysis to participants who have completed at least 9^{th} grade.

WJ Response:

The following code chunk generates Q-Q plots for the cases within each group (each education level) to assess their normality:

```
ggplot(data = clean_data,
       aes(sample = BMI, group=Education, colour=Education)) +
  geom_line(stat = "qq", )
```



While there is some deviation from a line-like share towards the ends of the plot, the middle values seem to approximate a simple sloped line. Given this, we can assume that the data within each group is approximately normal.

Next, we can check the variances of each group, which is completed in the cell below:

```
clean_data %>%
  group_by(Education) %>%
    summarize(group_variance = var(BMI))
```

```
## # A tibble: 5 x 2
##   Education      group_variance
##   <fct>                   <dbl>
## 1 8th Grade                16.7
## 2 9 - 11th Grade           72.9
## 3 High School              44.0
## 4 Some College             46.7
## 5 College Grad              42.2
```

The variances do reveal one potential point of concern. When comparing the ratios of the highest to lowest variances (the 8th grade and 9-11th grade groups, respectively) we get a factor of about 4.3. This is likely too large to assume that each group has equal variance. As such, is it advisable to remove this group from the analysis. Doing so means that the new lowest variance is from the College Grad group, and that the ratio between the highest and lowest variances is a much more acceptable 1.7. With this value, we can assume that the variances in each group are equal. The code cell below removes 8th grade group from our data frame:

```
clean_data <- clean_data %>%
  filter(Education != '8th Grade')
```

---

6. Conduct a hypothesis test to address the question of interest. Let $\alpha = 0.05$. Summarize the conclusions.

WJ Response:

Given the following:

$$
\begin{aligned}
\mu_1 &= \text{Mean BMI of 9-11th grade participants}\\
\mu_2 &= \text{Mean BMI of High School participants}\\
\mu_3 &= \text{Mean BMI of participants who have completed some college}\\
\mu_4 &= \text{Mean BMI of College Grad participants}
\end{aligned}
$$

we can define the following null and alternative hypotheses:

$$
\begin{aligned}
H_0 :&\mu_1 = \mu_2 = \mu_3 = \mu 4\\
H_1 :&\exists i, j \in [1, 4] \mid \mu_i \neq \mu_j
\end{aligned}
$$

The above hypothesis test is carried out below using the `aov` function:

```
aov_mod <- aov(BMI ~ Education, data = clean_data)
summary(aov_mod)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Education     3    217   72.35   1.495  0.215
## Residuals   460  22258   48.39
```

Given an α value of 0.05 and the p value displayed above of 0.215, we can conclude from the results above that the ANOVA test did not garner a statistically significant result. Thus, we cannot reject the null hypothesis and conclude that education level does not have a statistically significant impact on BMI. Furthermore, this result means that there is no point to doing any pairwise comparisons.

Chick weights: comparing weight across feed supplements

Chicken farming is a multi-billion dollar industry, and any methods that increase the growth rate of young chicks can reduce consumer costs while increasing company profits. An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chicks. Newly hatched chicks were randomly allocated into groups, and each group was given a different feed supplement.

The `chickwts` dataset available in the `datasets` package contains the weight in grams of chicks at six weeks of age. For simplicity, this analysis will be limited to four types of feed supplements: linseed, meatmeal, soybean, and sunflower.

7. Run the following code to load the `chickwts` dataset and subset the data for the four feed supplements of interest.

WJ Response:

```
#load the data
library(datasets)
data("chickwts")
#subset the four feed supplements
keep = (chickwts$feed == "linseed" | chickwts$feed == "meatmeal" |
  chickwts$feed == "soybean" | chickwts$feed == "sunflower")
chickwts = chickwts[keep, ]
#eliminate unused levels
chickwts$feed <- droplevels(chickwts$feed)
```

---

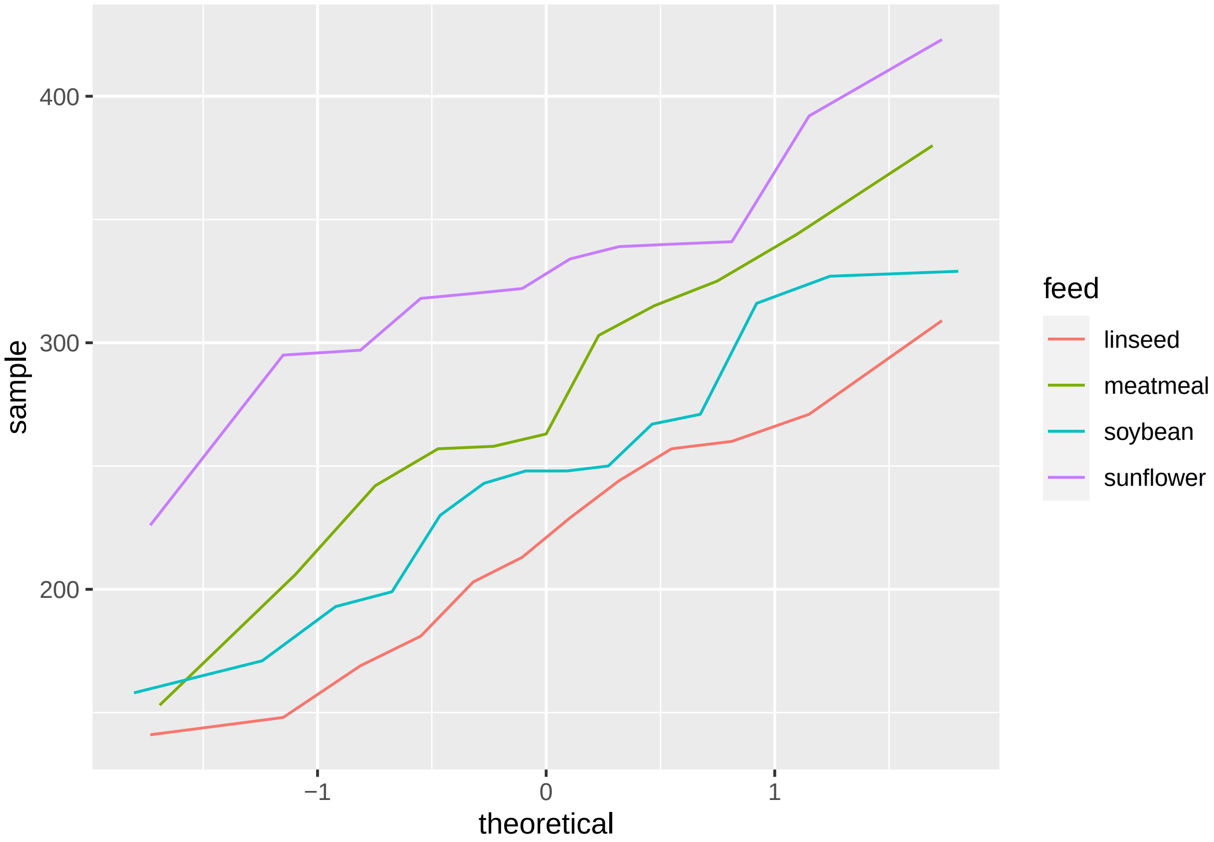
8. Analyze the data and report the results. Using language accessible to a non-statistician, discuss which feed supplement(s) is/are the most effective for increasing chick weight.

WJ Response:

~Checking assumptions~

First, we must check to make sure all the conditions for inference for an ANOVA test are satisfied. The cell below checks the normality of each of our groups groups:

```
ggplot(data = chickwts,
       aes(sample = weight, group=feed, colour=feed)) +
  geom_line(stat = "qq", )
```



As is clear in the plot above, the Q-Q line for each of the feeds seem to approximate a simple sloped line. Given this, we can assume that the data within each group is approximately normal.

Next, we check the variance of each group:

```
chickwts %>%
  group_by(feed) %>%
    summarize(group_variance = var(weight))
```

```
## # A tibble: 4 x 2
##   feed      group_variance
##   <fct>              <dbl>
## 1 linseed            2729.
## 2 meatmeal           4212.
## 3 soybean            2930.
## 4 sunflower          2385.
```

The proportion between the largest and lowest variance reported above is about 1.7. This factor is low enough to assume that the variances of the chick weights for each feed type are approximately the same.

~Running ANOVA Test~

Given the following:

$$\begin{aligned}
\mu_1 &= \text{Mean weight of chick that is fed linseed}\\
\mu_2 &= \text{Mean weight of chick that is fed meatmeal}\\
\mu_3 &= \text{Mean weight of chick that is fed soybean}\\
\mu_4 &= \text{Mean weight of chick that is fed sunflower}
\end{aligned}$$

we can define the following null and alternative hypotheses:

$$\begin{aligned} H_0 :& \mu_1 = \mu_2 = \mu_3 = \mu 4 \\ H_1 :& \exists i, j \in [1, 4] \mid \mu_i \neq \mu_j \end{aligned}$$

The following cell tests these hypotheses using the `aov` function:

```r
aov_mod <- aov(weight ~ feed, data = chickwts)
summary(aov_mod)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## feed         3  80941   26980   8.897 9.66e-05 ***
## Residuals   45 136460    3032                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting p-value in this case is extremely small (much smaller than our α value of 0.05), and thus we can conclude that there is at least 1 pair of feed types that have a statistically significant impact on average chick weights.

~Producing Paired Comparisons~

Now that we know there are differences in the impact of feed type on average chick weight, we can run paired comparisons to see between which groups those differences actually exist. However, we first need to adjust our level of significance from α to α^* using the Bonferroni correction. This is done in the cell below:

```r
num_groups <- length(unique(chickwts$feed))
alpha_i <- 0.05

K <- num_groups * (num_groups - 1) /2
alpha_f <- alpha_i / K
alpha_f
```

```
## [1] 0.008333333
```

Thus, the appropriate α level in this case is $\alpha^* = 0.0083333$.

The following cell runs t tests comparing the mean `weight` values for each pairing of `feed` type (once again, using the Bonferroni correction):

```r
pairwise.t.test(chickwts$weight, chickwts$feed, p.adj = 'bonf')
```

```
## 
##  Pairwise comparisons using t tests with pooled SD 
## 
## data:  chickwts$weight and chickwts$feed 
## 
##           linseed meatmeal soybean
## meatmeal  0.0898  -        -      
## soybean   1.0000  1.0000   -      
## sunflower 7.7e-05 0.1712   0.0025 
## 
## P value adjustment method: bonferroni
```

Using our α^* level of significance, we see that there are two pairings which produced statistically significant results: sunflower vs. linseed and sunflower vs. soybean.

~Conclusion and Recommendation~

To determine which feed is actually the best, we can look at the mean weight for each group:

```r
chickwts %>%
  group_by(feed) %>%
    summarise(feed_mean = mean(weight))
```

```
## # A tibble: 4 x 2
##   feed      feed_mean
##   <fct>         <dbl>
## 1 linseed        219.
## 2 meatmeal       277.
## 3 soybean        246.
## 4 sunflower      329.
```

Given that sunflower feed has the highest mean and that there was a statistically significant difference when comparing the average chick weight using of sunflower versus linseed or soybean feed, this is what I would recommend to a farmer to use if they want the fattest chicks after the first six weeks of their lives.