MODULE 2

ONE-WAY ANOVA

Module Objectives:

1. Understand the hypotheses tested with a one-way ANOVA.
2. Understand the models tested with a one-way ANOVA.
3. Identify the four assumptions of a one-way ANOVA.
4. Identify how to test the assumptions of a one-way ANOVA.
5. Understand why 2-sample t-tests cannot be used to make all pairwise comparisons.
6. Understand why multiple comparisons follow a significant one-way ANOVA.
7. Understand the distinction between individual-wise and experiment-wise error rates.
8. Understand the strengths and weaknesses of Tukey-Kramer HSD and Dunnett's multiple comparison procedures.
9. Understand what transformations are used for in a one-way ANOVA.
10. Understand how to use the trial-and-error method to choose a power transformation.
11. Present the results of a one-way ANOVA in an efficient, comprehensive, readable format.

A two-sample t-test is used specifically when the means of two independent populations are compared. Many realistic experiments and samples result in the comparison of means from more than two independent populations. For example, consider the following situations:

- Interest is in determining if the mean volume of white blood cells of Virginia opossums (*Didelphis virginiana*) differed by season in the same year (Woods and Hellgren 2003).
- Interest is in determining if the mean frequency of occurrence of badgers (*Meles meles*) in plots differs between plots at different locations (Virgos and Casanovas 1999).
- Interest is in testing for differences in the mean total richness of macroinvertebrates between the three zones of a river (Grubbs and Taylor 2004).
- Interest is in testing if the mean mass of porcupines (*Erithizon dorsatum*) differs among months of summer (Sweitzer and Berger 1993).
- Interest is in testing if the mean clutch size of spiders differs among three types of parental care categories (Simpson 1995).
- Interest is in determining if the mean age of harvested deer (*Odocoelius virginianus*) differs among deer harvested from Ashland, Bayfield, Douglas, and Iron counties.

In each of these situations, the mean of a quantitative variable (e.g., age, frequency of occurrence, total richness, or body mass) is compared among two or more populations of a single factor variable (e.g., county, locations, zones, or season). A two-sample t-test cannot be used in these situations because more than two groups are compared. A one-way analysis of variance (or **one-way ANOVA**) may be used in these situations. The theory and application of one-way ANOVAs are discussed in this module.[1]

◇ The two-sample t-test is used to determine if a significant difference exists between the means of two populations.

◇ A one-way analysis of variance (ANOVA) is used to determine if a significant difference exists among the means of more than two populations.

2.1 Analytical Foundation

The generic null hypothesis for a one-way ANOVA is

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_I$$

where I is the total number of groups identified by the factor variable. From this, it is evident that the one-way ANOVA is a direct extension of the two-sample t-test (see Section 1.1). The alternative hypothesis is complicated because not all pairs of means need differ for the null hypothesis to be rejected. Thus, the alternative hypothesis for a one-way ANOVA is "wordy" and is often written as

$$H_A : \text{“At least one pair of means is different”}$$

Thus, a rejection of the null hypothesis in favor of this alternative hypothesis is a statement that *some* difference in group means exists. It does not clearly indicate which group means differ. Methods to identify which group means differ are in Section 2.4.

[1] This presentation depends heavily on the foundational material in Module 1.

The simple and full models for the one-way ANOVA are the same as those for the two-sample t-test, except to note that there are $I > 2$ means in the full model (Figure 2.1). Thus, the total, within, and among SS are computed using the same formulas – i.e., Equations (1.3.2), (1.3.3), and (1.3.5) – except to note that $I > 2$. The degrees-of-freedom are computed similarly – i.e., $df_{Within} = n - I$ and $df_{Among} = I - 1$. The MS, F, and p-value are also computed the same.[2]

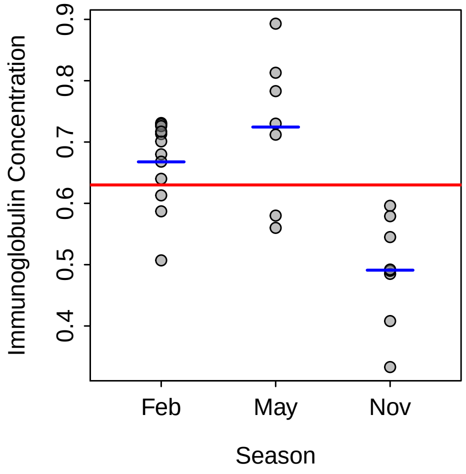

Figure 2.1. Immunoglobulin concentrations in Virginia opossums sampled from different seasons. The red horizontal line represents the simple model of a grand mean for both groups. The three blue horizontal lines represent the full model of separate means for each group.

◇ **The numerator df in a one-way ANOVA are $I - 1$.**

◇ **The denominator df in a one-way ANOVA are $n - I$.**

2.1.1 One-way ANOVA in R

Data Format

As with a two-sample t-test, the data for a one-way ANOVA must be stacked (Section 1.1.1). For example, the **Opposums.csv** (view, download, meta) file contains the immunoglobulin concentration levels measured on Virginia opossums sampled from three different seasons. The structure of the data file indicates two variables – *imm*, the immunoglobulin concentration levels, and *season*, the season the opossum was sampled – for 27 opossums. In addition, the structure indicates that the *season* variable is a factor with three levels. The display of several lines of the file shows the stacked nature of the data.

```
> opp <- read.csv("Opposums.csv")
```

```
> str(opp)
'data.frame': 27 obs. of  2 variables:
 $ imm   : num  0.64 0.68 0.731 0.587 0.668 0.613 0.713 0.701 0.729 0.726 ...
 $ season: Factor w/ 3 levels "feb","may","nov": 1 1 1 1 1 1 1 1 1 1 ...
```

[2] The MS, F, and p-value are computed the same in nearly every ANOVA table encountered in this class.

```
> headtail(opp)
     imm season
1  0.640    feb
2  0.680    feb
3  0.731    feb
25 0.490    nov
26 0.333    nov
27 0.492    nov
```

Recall that the levels of a factor variable are ordered alphabetically by default. In this instance, the alphabetical ordering is acceptable; i.e., “feb”, “may”, and “nov” is both the alphabetic and natural order for these levels. Changing the order of the levels was described in Section 1.1.3.

Fitting Model & Results

A one-way ANOVA model is fit with `lm()` exactly as described for a 2-sample t-test in Section 1.1.3.[3] For example, the very small p-value ($p = 0.0001$) below indicates that the full model of a separate mean for each group fits the data “better” than the simple model of one common mean. Thus, there is a significant difference in mean immunoglobulin level between at least one pair of the three seasons.

```
> opp.lm <- lm(imm˜season,data=opp)
> anova(opp.lm)
Analysis of Variance Table

Response: imm
          Df  Sum Sq  Mean Sq F value    Pr(>F)
season     2 0.23401 0.117005  14.449 7.609e-05
Residuals 24 0.19435 0.008098
```

The natural reaction at this point is to ask “Which means are different?”. This question will be answered more completely in Section 2.4. However, giving the saved linear model object to `fitPlot()` will produce a graphic to visually compare group means (Figure 2.2).

```
> fitPlot(opp.lm,xlab="Season",ylab="Immunoglobulin Concentration")
```

2.2 Assumptions

A one-way ANOVA has the same assumptions as a two-sample t-test. The four assumptions are

1. independence of individuals within and among groups,
2. equal variances among groups,
3. normality of residuals within each group, and
4. no outliers

[3] The `aov()` function can also be used. However, `aov()` calls `lm()` to make the calculations. For this reason, and the fact that `lm()` is more general and can be used for a wider variety of situations, only `lm()` is discussed here.

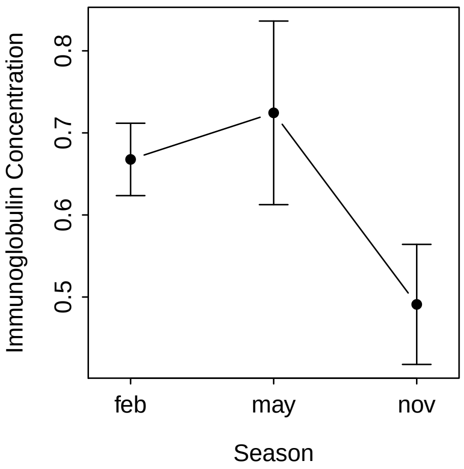

Figure 2.2. Mean (with 95% CI) immunoglobulin concentrations in Virginia opossums sampled from different seasons.

◇ The one-way ANOVA has four assumptions: independence among individuals, equal variances among groups, normality within groups, and no outliers.

It is critical to the proper analysis and interpretation of one-way ANOVAs that the individuals are independent both within and among groups. In other words, there must be no connections between the individuals within a group or between individuals among groups. Examples of a lack of independence include applying multiple treatments to the same individual (e.g., treatment A in week 1, treatment B in week 2, etc.), having all related individuals within the same group (e.g., all siblings are in the same group), or having individuals that are not separated in space and time (e.g., the first four individuals receive treatment A, the second four individuals receive treatment B, etc., or four clustered individuals are in group A, four other clustered individuals are in group B, etc.). Violations of this assumption are usually detected by careful consideration of the design of the data collection. Violations that are discovered after the data are collected cannot be corrected and the data have to be analyzed with techniques specific to dependent data. In other words, designing data collections with independence among individuals is critical and needs to be ascertained before the data are collected.

◇ Independence of individuals is a critical assumption of one-way ANOVAs. Violations of this assumption cannot be corrected.

The variances among groups must be equal because the estimate of MS_{Within} is based on a pooling of estimates from the individual groups. In other words, if the variances among each group are equal, then the MS within each group is an estimate of the overall MS_{Within}. In this instance, combining the values from each group provides a robust estimate of the overall variance within groups.

The assumption of equal variances can be tested with Levene's homogeneity of variances test.[4] The hypotheses tested by Levene's test are

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2 \\ H_A &: \text{“At least one pair of variances is different”} \end{aligned}$$

Thus, a p-value less than α means that the variances are not equal and the assumption of the one-way ANOVA has not been met.[5]

In certain instances,[6] Levene's test is not practical for examining the equality of variances. In these instances, the equality of variances may be visually examined with a boxplot of full model residuals by group. If the "boxes" on this boxplot are not roughtly the same, then the equal variances assumption may be violated. This boxplot should only be used if Levene's test cannot be used to test for equal variances.

◇ **Equal variances among groups is a critical assumption of a one-way ANOVA. Violations of this assumption should be corrected.**

◇ **The equal variance assumption is tested with Levene's test. P-values less than α indicate that the variances are not equal and the assumption was violated.**

The normality of residuals WITHIN each group is difficult to test because there may be (1) many groups being considered or (2) relatively few individuals in each group. Because most linear models are robust to slight departures from normality, it is often assumed that if the full model residuals are approximately normally distributed, then the residuals within each group are also normally distributed. Thus, the normality assumption for one-way ANOVA reduces to examining the normality of full model residuals together (i.e., not separated by groups).

The normality of residuals may be tested with the Anderson-Darling Normality Test.[7] In this instance, the hypotheses tested by an Anderson-Darling test are

$$\begin{aligned} H_0 &: \text{“Residuals are normally distributed”} \\ H_A &: \text{“Residuals are not normally distributed”} \end{aligned}$$

An Anderson-Darling p-value greater than α indicates that the residuals appear to be normally distributed and the normality assumptions is met. An Anderson-Darling p-value less than α suggests that the normality assumption has been violated.

◇ **The normality assumption is tested with the Anderson-Darling test of the full model residuals. P-values less than α indicate that the residuals are not normally distributed and the normality assumption was violated.**

[4] There are a wide variety of statistical tests for examining equality of variances. We will use the Levene's test in this class because it is common in the literature and simple to implement in most statistical software packages.

[5] Methods for "working around" this assumption are discussed in Section 2.6.

[6] For example, if there is a large number of groups with a small number of individuals each or if only one individual per block-treatment combination is used.

[7] There are also a wide variety of normality tests. Some authors even argue against the use of hypothesis tests for testing normality and suggest the use of graphical methods instead. For simplicity, the Anderson-Darling normality test will be used throughout this book.

As mentioned before, the one-way ANOVA is robust to slight departures from normality within groups. Some authors argue that a one-way ANOVA can still be used if the residuals from the one-way ANOVA fit are, at least, not strongly skewed and the sample size is moderately large. Thus, if the Anderson-Darling normality test suggests non-normality in the residuals, one should construct a histogram of the residuals to determine if they are not strongly skewed. If the residuals are strongly skewed, then the methods of Section 2.6 should be considered. If the residuals are only slightly skewed and the other assumptions have been met, then one can proceed relatively confidently with a one-way ANOVA.

◇ **A one-way ANOVA is robust to slight violations of the normality assumption. Severe violations of this assumption should be corrected.**

The one-way ANOVA is very sensitive to outliers. Outliers should be corrected if possible (usually if there is a data transcription or entry problem). The outlier should be deleted if it is determined that the outlier is clearly in error or is not part of the population of interest. If the outlier is not corrected or deleted, then the relative effect of the outlier on the analysis should be determined by completing the analysis with and without the outlier present. Any differences in results or interpretations due to the presence of the outlier should be clearly explained to the reader.

◇ **A one-way ANOVA is very sensitive to outliers.**

◇ **Outliers that are obvious errors should be fixed or deleted. The effect of outliers that are not errors should be assessed by completing the one-way ANOVA with and without the outlier in the data set.**

Outliers may be detected by visual examination of a residual plot. In addition, potential outliers can be more objectively detected with Studentized residuals, which is a residual divided by the standard deviation of the residual. Because residuals have a mean of zero, this calculation essentially computes how many standard deviations an individual residual is from the group mean. Becaus this is the standard definition of a t test statistic, Studentized residuals have the property of following a t distribution with $n - I$ df.

One problem with Studentized residuals is that the standard deviation of the residual is inflated if the individual is indeed an outlier. One method of correcting this problem is to compute the standard deviation of the residual with that residual removed from the data. This standard deviation is called the “leave-one-out” standard deviation and is common practice for many calculations aimed at finding potential outliers. A Studentized residual computed with the “leave-one-out” standard deviation is called an externally Studentized residual.[8] Externally Studentized residuals will be used exclusively in this book and will simply be called Studentized residuals.

The main advantage of Studentized residuals is that their distribution is well known – i.e., they follow a t distribution with degrees-of-freedom equal to $df_{Within} - 1$ or $n - I - 1$.[9] This allows construction of a hypothesis test to determine whether an individual can be considered to be a significant outlier or not. The p-value for this hypothesis test is calculated by converting the Studentized residual to a two-tailed p-value using a t distribution.

This method of testing for outliers is “dangerous” because the researcher will “sort through” all of the residuals to focus on the most extreme residual. So, in essence, the researcher constructs n hypothesis

[8] Some authors call these jackknife residuals.

[9] The extra one is subtracted because of the “leave-one-out” practice.

tests, but only focuses on one. This type of "testing" leads to a difficulty called the "problem of multiple comparisons", which is discussed in much more detail in Section 2.4. The multiple comparisons problem can be conservatively corrected with a Bonferroni correction, which constructs an adjusted p-value by multiplying the original p-value by the number of comparisons made (in this case n). If the Bonferroni adjusted p-value for the most extreme residual is less than α, then that individual is considered to be a significant outlier and should be flagged for further inspection as described above.

2.2.1 Assumption Checking in R

Equal Variances

The equal variances assumptions is tested with `levenesTest()` which requires a model formula[10] or an object from `lm` as its only argument. These results indicate that the variances among the three seasons appear to be equal ($p = 0.37$). Thus, the equal variances assumption of the one-way ANOVA has been met.

```
> levenesTest(opp.lm)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  1.0344 0.3708
      24
```

The equal variances assumption can also be visually assessed with a residual boxplot. The residual plot is constructed with `residPlot()` using the `lm()` object.[11] The residual boxplot shown in Figure 2.3-Left indicates approximately equal variances among the three groups because the "box" heights among seasons are "roughly" equal. Again, note that the Levene's test result is the definitive answer about equal variances in this instance.

```
> residPlot(opp.lm)
```

Normality

The Anderson-Darling normality test is performed by providing the vector of residuals from the saved `lm()` object to `adTest()`, as demonstrated below. The resulting p-value ($p = 0.0609$) indicates that there is only weak evidence that the residuals are not normally distributed.

```
> adTest(opp.lm$residuals)
Anderson-Darling normality test with x
A = 0.697, p-value = 0.0609
```

With weak evidence it is a good idea to construct a histogram of residuals to determine if there is any indication for strong skewness or, more likely, an outlier. The histogram of residuals was constructed with `residPlot()` above and is in Figure 2.3-Right. This histogram indicates a slight left-skewness in the residuals, but no strong evidence for an extreme skew or any outliers. The normality assumption can reasonably be said to have been met based on the results of the Anderson-Darling test and this histogram.

[10] As illustrated for the two-sample t-test.

[11] Raw residuals are used b default in `residPlot()`. Studentized residuals may be used by icluding `type="standardized"` in `residPlot()`.

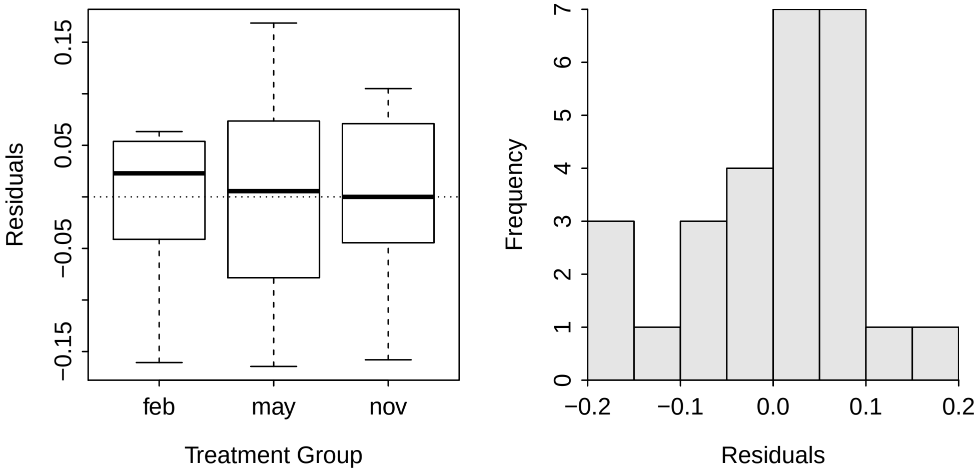

Figure 2.3. Residual plot (Left) and histogram of residuals (Right) for the one-way ANOVA of immunoglobulin concentrations in Virginia opossums sampled from different seasons.

Outliers

The Bonferroni adjusted p-value for the most extreme Studentized residual is computed with `outlierTest()`, which requires the saved `lm()` object as its only argument. With these data, individual 19 had an absolute value of the Studentized residual of 2.175 and a Bonferroni-adjusted $p > 1$.[12] This adjusted p-value is much greater than α and, thus, there is no indication of an outlier in these data.

```
> outlierTest(opp.lm)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
19 2.174897           0.040172           NA
```

Note that significant outliers, as identified with `outlierTest()`, will be marked with the observation number on the default residual plot from `residPlot()`.

2.3 Example Analyses I

2.3.1 Tomatoes-Nematodes I

Introduction

Nematodes are microscopic worms found in soil that may negatively affect the growth of plants through their trophic dynamics. Tomatoes are a commercially important plant species that may be negatively affected by high densities of nematodes in culture situations.

[12] Note that R will not show p-values greater than 1 and returns an `NA` instead. There is also a note at the beginning of this output that shows that no Bonferroni p-value is < 1.

A science fair student designed an experiment to determine the effect of increased densities of nematodes on the growth of tomato seedlings (i.e., an indicator of plant health). The student hypothesized that nematodes would negatively affect the growth of tomato seedlings – i.e., growth of seedlings would be lower at higher nematode densities. The statistical hypotheses to be examined were

$$H_0 : \mu_0 = \mu_{1000} = \mu_{5000} = \mu_{10000}$$
$$H_A : \text{“At least one pair of means is different”}$$

where the subscripts identify densities of nematodes (see below).

Data Collection

The student had 16 pots of a homogeneous soil type in which he “stocked” a known density of nematodes. The densities of nematodes used were 0, 1000, 5000, or 10000 nematodes per pot. The density of nematodes to be stocked in each pot was randomly assigned. After stocking the pots with nematodes, tomato seedlings, which had been selected to be as nearly identical in size and health as possible, were transplanted into each pot. The exact pot that a seedling was transplanted into was again randomly selected. Each pot was placed under a growing light in the same laboratory and allowed to grow for six weeks. Watering regimes and any other handling necessary during the six weeks was kept the same as possible among the pots. After six weeks, the plants were removed from the growing conditions and the growth of the seedling (in cm) from the beginning of the experiment was recorded.

Exploratory Data Analysis and Assumption Checking

It appears that tomato seedling growth may differ among nematode densities, with growth apparently suppressed at the two highest densities (Figure 2.4). The dispersion among individuals appears to be similar among the four groups (Figure 2.4).

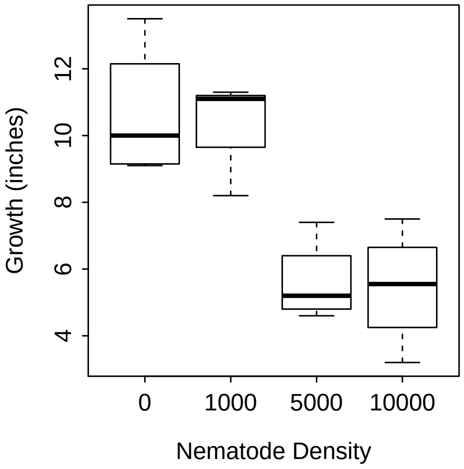

Figure 2.4. Boxplot of tomato seedling growth at each nematode density.

Individuals appear to be independent in this experiment because there does not appear to be any connection among pots either within (this assumes that the pots were randomly placed in the laboratory) or among treatments. Variances among the treatments appear to be approximately constant (Levene's $p = 0.8072$; Figure 2.5-Left). Anderson-Darling normality tests on the residuals of the initial model fit indicates that the residuals are normally distributed ($p = 0.9268$) and the histogram does not indicate any major problems (Figure 2.5-Right). There does not appear to be any major outliers in the data (Figure 2.4). One individual in the 0 nematode group had a Studentized residuals of 2.3011 but with a Bonferroni p-value of 0.6712, it was not considered to be an outlier. The analysis will proceed because the major assumptions of the one-way ANOVA have been met.

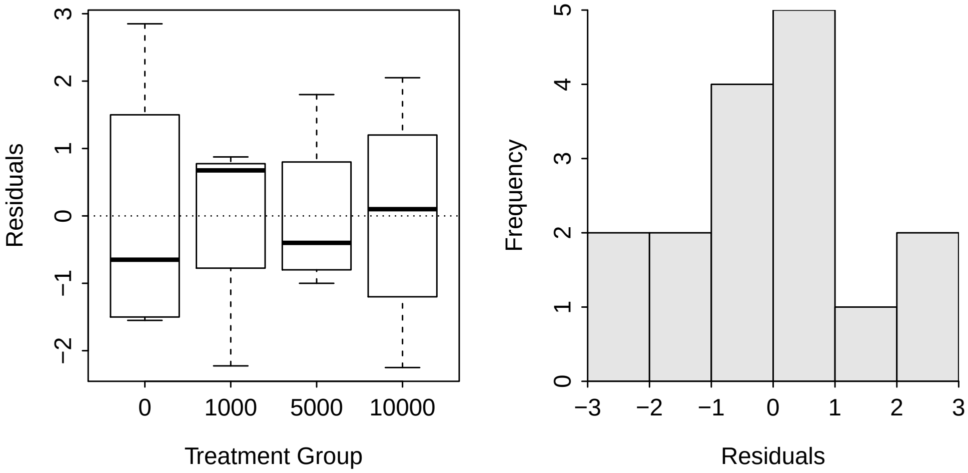

Figure 2.5. Boxplot (Left) and histogram (Right) of residuals from the initial fit of the one-way ANOVA model to the tomato seedling growth at each nematode density.

Results

There appears to be a significant difference in mean tomato seedling growth among the four treatments ($p = 0.0006$; Table 2.1). The plot of each treatment mean with 95% confidence intervals indicates that the mean growth at the two lowest nematode densities probably are not different and the mean growth at the two highest nematode densities probably are not different, but the mean growth at the two lowest nematode densities are different from the two highest nematode densities (Figure 2.6).[13]

Table 2.1. ANOVA results for tomato seedling growth at four nematode densities.

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
density    3 100.647  33.549   12.08 0.0006163
Residuals 12  33.327   2.777
```

Conclusion

The student's hypothesis was generally supported; however, it does not appear that tomato seedling growth is negatively affected for all increases in nematode density. For example, seedling growth declined for an

[13] Objective methods for determining which treatment means are significantly different are discussed in Section 2.4.

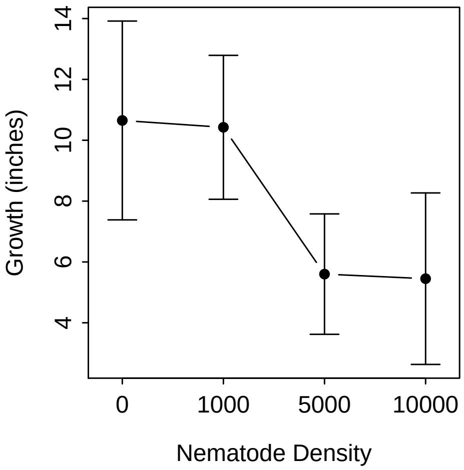

Figure 2.6. Mean tomato seedling growth with 95% confidence interval at each nematode density from the fit of the one-way ANOVA model.

increase in nematode density from 1000 to 5000 per pot but not for increases from 0 to 1000 nematodes per pot or from 5000 to 10000 nematodes per pot.

It can be concluded that the different nematode densities caused the differences in tomato seedling growth because the individual seedlings were randomly allocated to treatment groups and all other variables were controlled. However, the inferences cannot be extended to a general population of tomatoes because the 16 seedlings used in the experiment were not randomly chosen from the population of seedlings.

From these results, the experimenter might want to re-run the experiment for densities between 1000 and 5000 nematodes per pot in an attempt to find a "critical" nematode density below which there is very little affect on growth and above which there is a significant negative affect on growth.

Appendix – R Commands

```
TomatoNematode <- read.csv("TomatoNematode.csv")
TomatoNematode$density <- factor(TomatoNematode$density)
boxplot(growth~density,data=TomatoNematode,xlab="Nematode Density",ylab="Growth (inches)")
tn.lm <- lm(growth~density,data=TomatoNematode)
levenesTest(tn.lm)
adTest(tn.lm$residuals)
residPlot(tn.lm)
outlierTest(tn.lm)
anova(tn.lm)
fitPlot(tn.lm,xlab="Nematode Density",ylab="Growth (inches)")
```

2.3.2 Moose-Pines I

Introduction

The availability of resources for growth is believed to have a substantial impact on the chemical defense of plants against herbivores. However, the means by which resource availability affects different plant traits, and the way in which these factors in turn affect diet selection by herbivores are not well understood. Edenius (1993) addressed the relation between plant biomass, morphology, and tissue nutritional quality and browsing by moose (*Alces alces*) on Scots pine (*Pinus sylvestris*).

Data Collection

In one part of this study, Edenius examined the effect of three different experimental treatments related to nutrient and light availability on various characteristics of the Scots pine. In this example, the characteristic of the Scots pine that will be examined is tree height (measured in cm). The four treatments were labeled as "Fertilized", "Clipped", "Shaded", and "Control." In the fertilized treatment, 60 g of nitrogen (ammonium nitrate) was applied to the soil within a 2-m radius of each tree at the beginning of the growing season. In the clipped treatment, all shoots produced in the previous growing year were removed. In the shaded treatment, the top- and lateral-most branches were covered with a shade cloth that reduced the light intensity by 50% in the 400-700 nm wavelengths. Finally, a fourth group of trees were maintained without any manipulation as a control.

A total of 140 unbrowsed trees that were approximately 1.4 m in height were specifically selected for use in the experiment. The trees were randomly allocated to the four treatments such that each group had 35 trees in it. Selected trees were separated by at least 5 m to avoid interference among individual trees and, thus, treatment groups. Trees were allowed to grow for one full growing season and then were measured for height. Edenius, wanted to determine if there was a significantly different mean height among the treatments. Thus,

$$\begin{aligned} H_0 &: \mu_{Fert} = \mu_{Clip} = \mu_{Shade} = \mu_{Control} \\ H_A &: \text{“At least one pair of means is different”} \end{aligned}$$

One-way ANOVA will be used to identify if any significant differences exist among the means of the treatment groups.

Exploratory Data Analysis and Assumption Checking

It appears that tree height differs among treatments, with the clipped group being substantially smaller than the other three groups (Table 2.2, Figure 2.7). There is also some indication that the variances might be different as the standard deviation for the "control" group appears to be substantially larger than the standard deviation for the "shaded" group (Table 2.2).

The random allocation of trees to the treatments and the realization that applying the treatment to any one tree has no affect on any other tree implies that there is independence both within a treatment and among treatments. Variances among the treatments may be non-constant (Levene's $p = 0.0492$). However, the residual plot (Figure 2.8-Left) does not indicate any extreme differences in variances and no transformation (see Section 2.6) appeared to correct this problem. The residuals from the initial model fit appear to be approximately normal (Anderson-Darling $p = 0.5427$). None of the individuals appeared to be significant outliers as the largest absolute value Studentized residual was -2.861 with a Bonferroni-adjusted p-value of 0.6852. Thus, the analysis will continue with a one-way ANOVA as the assumptions either appear to be met, are not grossly unmet, or no reasonable solution to the problems exists.

Table 2.2. Descriptive statistics for height of Scots pines in four treatment groups.

```
      treat  n     mean        sd   min     Q1 median     Q3   max
    Control 35 165.3514 11.392138 144.9 157.55  164.7 173.15 185.6
 Fertilized 35 170.8857  9.915352 145.9 165.35  171.6 175.15 192.8
    Clipped 35 131.9314  7.838846 117.1 126.10  131.0 138.55 150.0
     Shaded 35 163.9514  6.409156 150.7 159.25  162.8 169.25 175.3
```

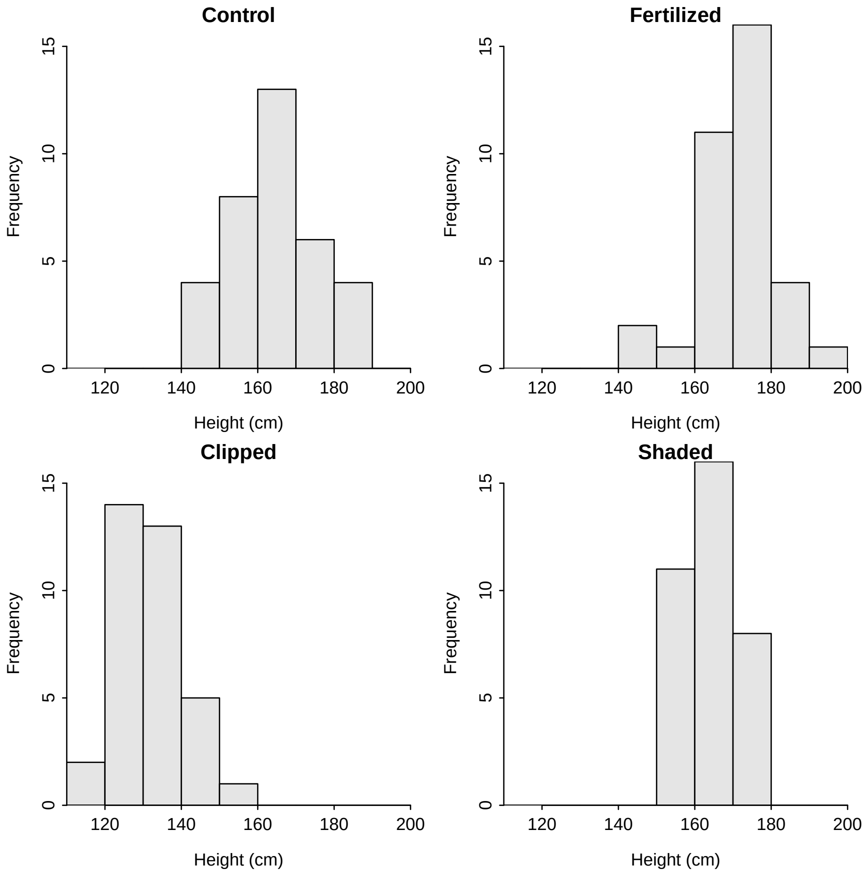

Figure 2.7. Histograms of Scots pine height for four treatment groups.

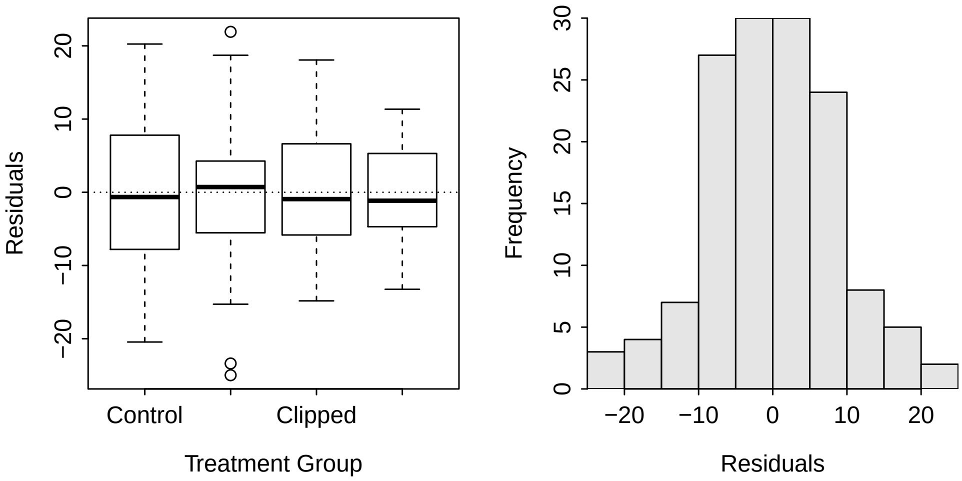

Figure 2.8. Boxplot (Left) and histogram (Right) of residuals from the initial fit of the one-way ANOVA model to the Scots pines heights at each treatment level.

Results

There appears to be a significant difference in mean tree growth among the four treatments ($p < 0.00005$; Table 2.3). Plots for each treatment group indicate that mean height for the clipped treatment is lower than the mean height in all other treatments, the mean height in the fertilized treatment may be greater than the mean height in all other treatments, and the mean heights in the shaded and control groups do not differ (Figure 2.9).

Table 2.3. ANOVA results for tree growth for four treatments.

```
           Df Sum Sq Mean Sq F value    Pr(>F)
treat       3  32728 10909.2  131.98 < 2.2e-16
Residuals 136  11241    82.7
```

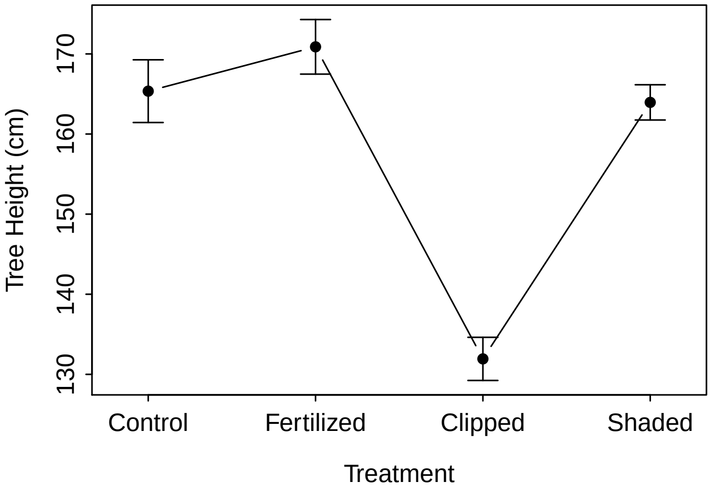

Figure 2.9. Mean Scots pine height with 95% confidence interval for each treatment group from the initial fit of the one-way ANOVA model.

Conclusion

The clipped treatment resulted in significantly lower growth of Scots pine. The fertilized treatment may have produced slightly taller trees than the control group.

It can be concluded that the different treatments caused the differences in tree growth because the individual trees were randomly allocated to treatment groups and all other variables were controlled. However, the inferences cannot be extended to a general population of trees because the 140 trees used in the experiment were not randomly chosen from the population of trees.

Appendix – R Commands

```
MooseBrowse <- read.csv("MooseBrowse.csv")
MooseBrowse$treat <- factor(MooseBrowse$treat,
   levels=c("Control","Fertilized","Clipped","Shaded"))
Summarize(height~treat,data=MooseBrowse)
hist(height~treat,data=MooseBrowse,xlab="Treatment",ylab="Tree Height (cm)")
mb.lm <- lm(height~treat,data=MooseBrowse)
levenesTest(mb.lm)
adTest(mb.lm$residuals)
residPlot(mb.lm)
outlierTest(mb.lm)
anova(mb.lm)
fitPlot(mb.lm,xlab="Treatment",ylab="Tree Height (cm)")
```

2.4 Multiple Comparisons

A significant result (i.e., reject H_0) in a one-way ANOVA indicates that the means of at least one pair of groups differ. At this time, it is not known whether all means are different, two means are equivalent but different from all other means, all means are equivalent except for one pair, or any other possible combination of equivalencies and differences. Thus, once a significant overall result is obtained in a one-way ANOVA, specific follow-up analyses are needed to identify which pairs of means are significantly different.

◇ **A one-way ANOVA only indicates that at least one pair of means differ. Follow-up analyses are required to specifically determine which pairs of means are different.**

2.4.1 The Problem

The most obvious solution for identifying which pairs of means are different is to perform multiple 2-sample t-tests on all pairs of groups. Unfortunately, this seemingly simple answer has at least two major difficulties. First, the number of 2-sample t-tests needed increases dramatically with increasing numbers of groups (Table 2.4). Second, the probability of incorrectly concluding that at least one pair of means differs when no pairs actually differ increases dramatically with increasing numbers of groups (Table 2.4). Of these two difficulties, the second is much more problematic and needs to be better understood.

Table 2.4. Relationship between the number of groups in an analysis, the number of pairs of means that would need to be tested and the experiment-wise error rate for two different rejection criteria.

Groups	Number of Pairs to Test	$\alpha = 0.05$	$\alpha = 0.10$
2	1	0.05	0.10
3	3	0.1426	0.2710
4	6	0.2649	0.4686
5	10	0.4013	0.6513
6	15	0.5367	0.7941

In any one comparison of two means the probability that one will incorrectly conclude that the means are different when they are actually not different is α. This incorrect conclusion is called a Type I error and α

is called the *individual-wise Type I error rate* because it relates to one individual comparison of a pair of means.

Δ **Individual-wise error rate**: The probability of a Type I error in a single comparison of two means. The individual-wise error rate is set at α.

◇ **A Type I error is rejecting the H_0 when the H_0 is actually true. In a two-sample t-test, a Type I error is concluding that the two means are significantly different when in fact they are not.**

If three pairs of means are simultaneously compared, as would happen if there were three groups (Table 2.4), then the probability that at least one decision for a pair of these means will be incorrect increases. The probability that **at least** one Type I error occurs in simultaneous comparisons is called the *experiment-wise error rate* because it involves all comparisons in the experiment at hand. It is very important that you notice the words *at least* in the previous sentences. In three comparisons, the incorrect conclusion could be for the first pair, the second pair, the third pair, the first and second pair, the first and third pair, the second and third pair, or all three pairs!! Because of this, the experiment-wise error rate is computed as the complement of the probability of no errors, or $1-(1-\alpha)^k$, where k is the number of paired comparisons to be made.

In Table 2.4, it is evident that, with $\alpha = 0.05$ and six treatments, the probability of concluding that at least one pair of means is different when there are no true differences among means is over 50%. In other words, it is nearly a coin flip that at least one error will be made in this situation. Six treatments is not a large set of groups to compare and this level of error is unacceptable and must be reduced.

Δ **Experiment-wise error rate**: The probability of at least one Type I error in a set of comparisons of two means. The experiment-wise error rate depends on the number of comparisons made and is calculated with $1-(1-\alpha)^k$, where k is the number of paired comparisons to be made.

◇ **The experiment-wise error rate increases dramatically with increasing numbers of treatment groups.**

2.4.2 Correction Methods

The statistical literature is full of various methods that have been designed to attempt to control the experiment-wise error rates. For simplicity, only three methods will be considered in this book. The Tukey-Kramer honestly significantly different (i.e., Tukey's HSD) method controls the experiment-wise error rate at a desired level (i.e., α) when the group sample sizes are the same and is slightly conservative if the group sample sizes are different. The Tukey's HSD method is the preferred multiple comparison method when all pairs of means are being compared. Dunnet's method is used to compare all groups to a specific control group. Dunnet's method does not make all pairwise comparisons, which increases statistical power compared to the Tukey's HSD method when comparing to a control group. Thus, even though both methods control the experiment-wise error rate, Dunnet's method should be used over the Tukey's HSD method when comparisons are made to a specific control group because the Dunnet's method will have higher statistical power (i.e., greater probability of identifying a true difference in means). The third method to control experimentwise error rates is discussed in Module 5.

◇ Use Tukey's HSD multiple comparison method when comparing all pairwise means.

◇ Use Dunnett's multiple comparison method when comparing all means to a single control group.

2.4.3 Multiple Comparisons in R

Tukey's HSD

Tukey's HSD procedure is applied using `glht()` from the `multcomp` package. This function requires a saved `lm()` object as its first argument and the `mcp()` function as its second argument. The factor to be considered must be "set equal" to `"Tukey"` in `mcp()` to obtain Tukey's HSD correction. The result of `glht()` should be assigned to an object that is then submitted to `summary()` to extract the adjusted p-values for all comparisons or `confint()` to find the confidence intervals for all pairs of differences. For example, the results for the immunoglobulin opossum data confirm that February and November and May and November are significantly different whereas February and May are not significantly different.[14] In fact, the mean immunoglobulin concentrations for opossums sampled in November appears to be between 0.074 and 0.279 lower than for opossums sampled in February and between 0.117 and 0.350 lower than for opossums sampled in May.[15]

```
> opp.mc <- glht(opp.lm,mcp(season="Tukey"))
> summary(opp.mc)

                Estimate Std. Error   t value      p value
may - feb = 0  0.0567619 0.04279830  1.326266 0.3939105145
nov - feb = 0 -0.1766667 0.04107416 -4.301163 0.0006563864
nov - may = 0 -0.2334286 0.04657372 -5.012023 0.0001051616
```

```
> confint(opp.mc)

            Estimate         lwr         upr
may - feb  0.0567619 -0.05001411  0.16353792
nov - feb -0.1766667 -0.27914120 -0.07419213
nov - may -0.2334286 -0.34962377 -0.11723337
```

Dunnett's method

Dunnett's procedure may also be applied using `glht()` with the factor variable in `mcp()` "set equal" to `"Dunnett"`. For example, the results below indicate that the mean immunoglobulin levels for opossums in May is not significantly different from opossums in February but the mean for opossums in November is significantly different from opossums in February.

[14] This result is "obtained" by comparing the adjusted p-values to α or by noting which confidence intervals for the differences in means do not contain zero.

[15] The results from `summary()` and `confint()` on a `glht()` object is overly verbose. Thus, the results have been condensed for printing in this book.

```
> opp.mc2 <- glht(opp.lm,mcp(season="Dunnett"))
> summary(opp.mc2)

                Estimate Std. Error   t value      p value
may - feb = 0  0.0567619 0.04279830  1.326266 0.3370801781
nov - feb = 0 -0.1766667 0.04107416 -4.301163 0.0004836763
```

```
> confint(opp.mc2)

            Estimate         lwr         upr
may - feb  0.0567619 -0.04437157  0.15789538
nov - feb -0.1766667 -0.27372596 -0.07960737
```

It is vitally important to note that all other groups will be compared to the group that is the first level in the factor variable. If the “base” group is not the first level of the factor variable, then the levels will need to be changed with `factor()` as shown in Section 1.1.3. Suppose, for example, that you wanted “may” to be the group that “feb” and “nov” would be compared to. The factor would then need to be re-leveled, a new linear model fit and saved, and this new model sent to `glht()`.

```
> opp$season1 <- factor(opp$season,levels=c("may","feb","nov"))
> opp.lm2 <- lm(imm~season1,data=opp)
> opp.mc3 <- glht(opp.lm2,mcp(season1="Dunnett"))
> summary(opp.mc3)

                Estimate Std. Error   t value      p value
feb - may = 0 -0.0567619 0.04279830 -1.326266 3.175589e-01
nov - may = 0 -0.2334286 0.04657372 -5.012023 7.787847e-05
```

Note that using the Dunnett’s procedure is unwarranted in this example – it is used here for the sole purpose of illustrating the method.

Graphing Significance Results

Multiple comparison results are often reported as ”significance letters” on a plot of group means (usually with corresponding confidence intervals). Significance letters are assigned such that group means with the same letter are considered statistically the same (i.e., insignificant) and group means with different letters are considered statistically different (i.e., significant). The Tukey HSD results for the opossum immunoglobulin data from above indicated that February and May should have the same letter (e.g., “a”) and November should have a different letter (e.g., “b”).

The plot of the group means is constructed with `fitPlot()` as illustrated previously. Significance letters are added to this plot with `addSigLetters()`. The first argument to `addSigLetters()` is the saved `lm()` object. The `lets=` argument is a character vector containing the letters to be placed next to each group mean, in the order that the group means are plotted. The `pos=` argument contains a numeric vector of positions that describe the position the letter should be placed relative to the point, with `1`=“below”, `2`=“left-of”, `3`=“above”, and `4`=“right-of”.[16] Finding “good” `pos` values may take some trial-and-error. Figure 2.10 was constructed with the code below.

[16] Note that the numbers are clockwise around the point beginning below the point.

```
> fitPlot(opp.lm,xlab="Season",ylab="Immunoglobulin level")
> addSigLetters(opp.lm,lets=c("a","a","b"),pos=c(2,4,4))
```

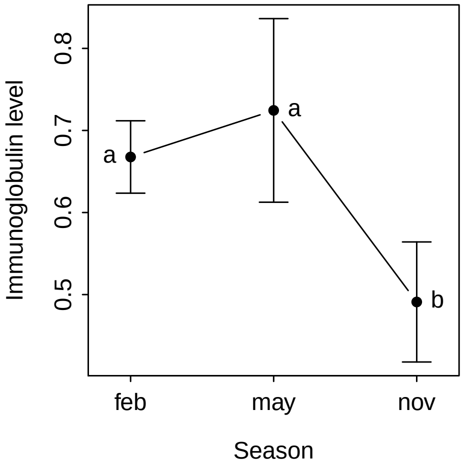

Figure 2.10. Mean tomato seedling growth with 95% confidence interval at each nematode density from the initial fit of the one-way ANOVA model. Means with the same letter are not significantly different.

2.5 Example Analyses II

2.5.1 Tomatoes-Nematodes II

In Section 2.3.1 the growth of tomato plants relative to the density of nematodes was examined. The one-way ANOVA results indicated that there was a significant difference in mean plant growth among the four densities of nematodes examined. Tukey's HSD procedure is used here to determine which pairs of groups means differ.

It appears that mean growth at densities 0 and 1000 are not significantly different, 0 and 5000 are different, 0 and 10000 are different, 1000 and 5000 are different, 1000 and 10000 are different, and 5000 and 10000 are not different (Table 2.5, Figure 2.11).[17]

Appendix – R commands

```
tn.mc <- glht(tn.lm, mcp(density="Tukey"))
summary(tn.mc)
confint(tn.mc)
fitPlot(tn.lm,xlab="Treatment",ylab="Tree Height (cm)")
addSigLetters(tn.lm,lets=c("a","a","b","b"),pos=c(2,4,2,4))
```

[17] Recall that two means are considered different if the adjusted p-value is less than α.

Table 2.5. Tukey's multiple comparisons for the Tomato - Nematode data.

```
                  Estimate Std. Error    t value     p value
1000 - 0 = 0        -0.225   1.178408 -0.1909355 0.997392445
5000 - 0 = 0        -5.050   1.178408 -4.2854421 0.004917151
10000 - 0 = 0       -5.200   1.178408 -4.4127324 0.004312913
5000 - 1000 = 0     -4.825   1.178408 -4.0945065 0.007116804
10000 - 1000 = 0    -4.975   1.178408 -4.2217969 0.005604022
10000 - 5000 = 0    -0.150   1.178408 -0.1272904 0.999219831
```

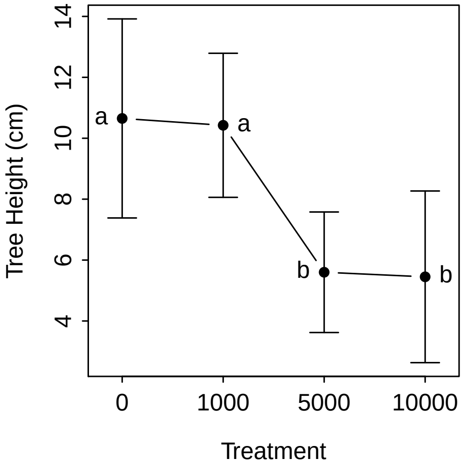

Figure 2.11. Plot of group means versus treatment levels with means that are statistically the same marked with the same letter.

2.5.2 Moose-Pines II

This example is a follow-up analysis to the second example in Section 2.3.2. Only those sections that would be modified to include multiple comparison results are shown here. Thus, a full analysis would be a combination of what was shown in Section 2.3.2 and what is shown here.

Data Collection

One-way ANOVA will be used to identify if significant differences exist among the means of the treatment groups. If a significant difference is identified, then Tukey's HSD method will be used to determine which pairs of treatment means are different.[18]

Results

There appears to be a significant difference in mean tree growth among the four treatments ($p < 0.00005$; Table 2.3). Trees in the clipped treatment are significantly shorter then trees in the other three treatments

[18] Dunnett's method is not used here, even though there is a control group, because interest is in comparing all pairs of treatments, not just all pairs of treatments with the control group.

(Table 2.6). The trees in the shaded treatment are shorter than trees in the fertilized treatment but statistically similar to trees in the control treatment (Table 2.6). Trees in the control treatment are statistically similar to trees in both the shaded and fertilized treatments (Table 2.6).[19] The results of this analysis are summarized in Figure 2.12.

Table 2.6. Tukey's adjusted confidence intervals for mean tree growth for four treatments. Note that the output was modified to save space.

```
                            Estimate Std. Error    t value     p value
Fertilized - Control = 0    5.534286   2.173279   2.546515 0.057524530
Clipped - Control = 0     -33.420000   2.173279 -15.377688 0.000000000
Shaded - Control = 0       -1.400000   2.173279  -0.644188 0.917410188
Clipped - Fertilized = 0 -38.954286   2.173279 -17.924202 0.000000000
Shaded - Fertilized = 0    -6.934286   2.173279  -3.190703 0.009126636
Shaded - Clipped = 0       32.020000   2.173279  14.733500 0.000000000
```

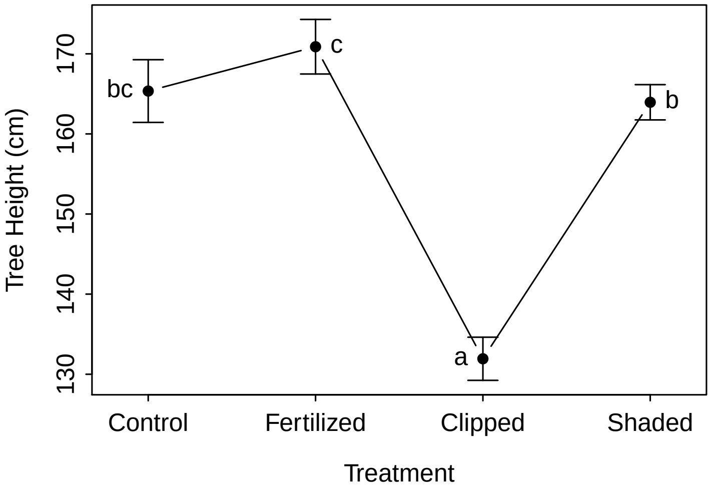

Figure 2.12. Plot of group means versus treatment levels with means that are statistically the same marked with the same letter.

Appendix – R commands

```
mb.mc <- glht(mb.lm, mcp(treat="Tukey"))
summary(mb.mc)
confint(mb.mc)
fitPlot(mb.lm,xlab="Treatment",ylab="Tree Height (cm)")
addSigLetters(mb.lm,lets=c("bc","c","a","b"),pos=c(2,4,2,4))
```

[19] This result for the shaded, control, and fertilized treatments is a fairly common occurrence - i.e., the middle of the ordered treatments is statistically similar to both the treatment just bigger and the treatment just smaller, but the two treatments on the ends are statistical different. So, sometimes the results lead to confusing but ultimately correct statements such as — "the control treatment is equal to both the shaded and fertilized treatments but the shaded and fertilized treatments are different."

2.6 Transformations

If the assumptions of a one-way ANOVA are violated, then the results of the one-way ANOVA are inappropriate. Fortunately, if the equality of variances or normality assumptions are violated, then corrective measures can usually be taken so that appropriate results can be obtained. The most common corrective measure is to transform the response variable to a scale where the variances among treatment groups are equal and the individuals within treatment groups are normally distributed.

Besides the obvious reason related to assumption violations, Fox (1997) gave four arguments for why data that is skewed or shows a non-constant variance should be transformed:

- Highly skewed distributions are difficult to examine because most of the observations are confined to a small part of the range of the data.
- Apparently outlying individuals in the direction of the skew are brought in towards the main body of the data when the distribution is made more symmetric. In contrast, unusual values in the direction opposite to the skew can be hidden prior to transforming the data.
- Linear models summarize distributions based on means. The mean of a skewed distribution is not, however, a good summary of its center.
- When a variable has very different degrees of variation in different groups, it becomes difficult to examine the data and to compare differences in levels across the groups.

The identification of an appropriate transformation and the understanding of the resultant output is the focus of this section.

◇ **If the assumptions of a one-way ANOVA are not met, then the data may be transformed to a scale where the assumptions are met.**

2.6.1 Families of Transformations

There are two major families of transformations – power and special transformations. Special transformations are generally identified based on the type of data to be transformed. Certain special transformations are common in particular fields of study and are generally well-known to scientists in those fields. An example that crosses many fields is the transformation of proportions or percentages data by using the arcsine square-root function ($sin^{-1}(\sqrt{Y})$). The effect of this transformation on right-skewed proportions data is illustrated in Figure 2.13. The histogram for the values on the original scale is shown upside-down in the lower portion of this figure, the transforming function is shown in the middle, and the resultant transformed data is shown sideways on the left. The "spreading out" and "compressing" can be visualized by drawing a line up from the original scale until the transforming function is met and then moving to the left until the transformed scale is met. This should be tried for a variety of values on the original scale to "feel" how the $sin^{-1}(\sqrt{Y})$ transformation spreads out small values and compresses large values.

With power transformations, the response variable is transformed by raising it to a particular power, λ, i.e., Y^{λ} (Table 2.7). Each of the power transformations shown in Table 2.7 tends to "spread out" relatively small values and "draw in" or "compress" relatively large values in a distribution. This process is illustrated in Figure 2.14 for a natural log transformation.[20]

[20] Try a few values as was done with the $sin^{-1}(\sqrt{Y})$ transformation function in Figure 2.13.

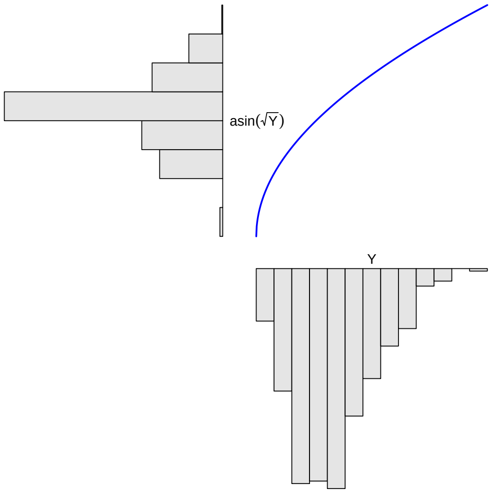

Figure 2.13. Demonstration of the result (left) from applying the arcsine square-root transformation function (blue line) to right-skewed original values (lower).

Table 2.7. List of common power transformations in ANOVAs.

Power	Transformation		Power	Transformation	
λ	Name	Formula	λ	Name	Formula
1	Original Scale	$Y^1 = Y$	0	Natural Log	$log(Y)$
0.5	Square Root	$Y^{0.5} = \sqrt{Y}$	-0.5	Inverse Root	$Y^{-0.5} = \frac{1}{\sqrt{Y}}$
0.33	Cube Root	$Y^{0.33} = \sqrt[3]{Y}$	-1	Inverse	$Y^{-1} = \frac{1}{Y}$
0.25	Fourth Root	$Y^{0.25} = \sqrt[4]{Y}$			

The common transformations listed in Table 2.7 are ordered from least to most powerful moving down the first column and then down the second. In other words, the transformations are listed in order from the transformations that "spread out" the small values the least to those that "spread out" the small values the most. This ordering can be seen by comparing the transforming functions in Figure 2.15. Alternatively, the transformations are ordered from those that "normalize" mildly skewed data to those that "normalize" strongly skewed data.

You should also note that it is possible to "combine" one of the common powers with the inverse transformation to create a larger array of inverse transformations. For example, a λ of -0.5 could be considered an inverse square-root transformation. These types of transformations are common but less common than those listed in Table 2.7.

Finding A Power Transformation

Power transformations require non-negative and non-zero data. Violations of this restriction can be rectified by adding an amount to all values of the response variable such that the all values become positive. In

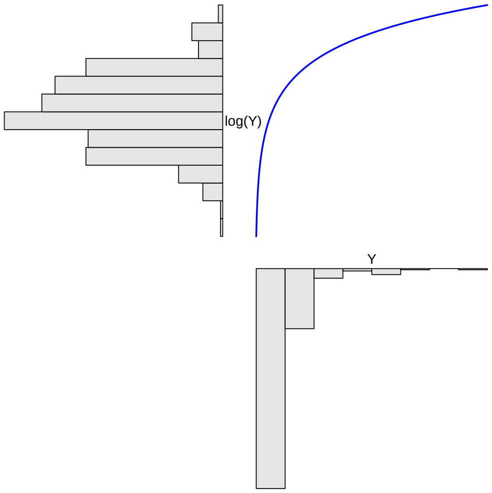

Figure 2.14. Demonstration of the result (upper-left) from applying the natural log transformation function (blue line in upper-right) to right-skewed original values (lower-right).

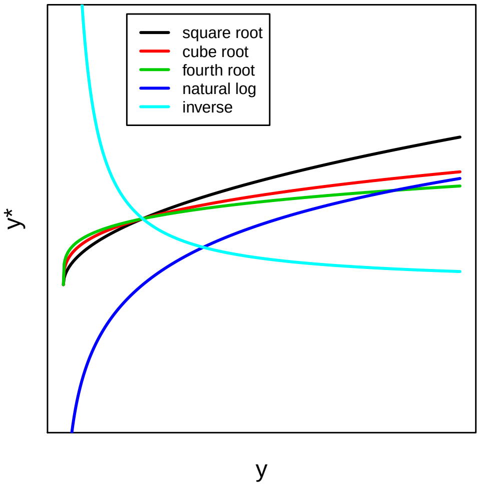

Figure 2.15. Most common power transformation functions.

addition, power transformations are not effective if the range of values of the response variable is narrow.[21] In general, Fox (1997) suggests that if the ratio of the maximum to minimum value of the response variable is less than five, then you should consider subtracting a value just smaller than the minimum value to shift the entire distribution of values closer to zero. These values used to "shift" the distribution are commonly called *start* values.

◇ **A positive value should be added to each value of the response variable if a power transformation is to be used and the response variable contains zeroes or negative values.**

[21] In effect, the power transformation is basically linear over short ranges and, thus, is not effective.

◇ A negative value may be added to each value of the response variable if a power transformation is to be used and the ratio of maximum to minimum value of the response variable is less than five.

There are several methods for identifying the power transformation that is most likely to correct problems with assumptions for a one-way ANOVA.[22] One simple method is trial-and-error – i.e., trying various powers until one is found where the assumptions of the model are most closely met. This trial-and-error method is tedious but is made more efficient with the use of `transChooser()`. This function requires a saved `lm()` object from the initial one-way ANOVA model fit as its first argument. In addition, a start value can be included with the `starty=` argument. This function then produces a histogram of the residuals of the model and a boxplot of the residuals separated for each group. A slider can then be used to "try" various values of λ with the histogram and boxplot being updated as λ is changed. One can try various values of λ until a value is found that provides approximately normal residuals and approximately equal variances. When actually transforming the variable it is best to choose one of the common values of λ from Table 2.7 that is closest to the value found by trial-and-error.

Another option for choosing a possible power transformation is to rely on theory related to the response variable. As with the special transformations discussed above, power transformations based on theory will generally be well-known by the scientists within a particular field. However, for example, it is common to transform response variables that are areas by taking the square root and those that are volumes with the cube root. In addition, discrete counts are often transformed with the square root.

◇ Sometimes transformations may be chosen based on known theory regarding the response variable.

2.6.2 Creating Transformed Variables in R

A power transformation can be carried out by raising the variable to the given power with the ˆ symbol. A square-root transformation can also be used by including the variable in `sqrt()`. A natural log transformation is accomplished by including the variable in `log()`. The arcsine square root transformation requires the combined efforts of `asin()` and `sqrt()`. The results of the transforming function should be assigned to a new object. The following are examples of creating a transformed variable from a generic variable called Y:

```
sqrt.y <- Y^(1/2)        # square-root
sqrt.y <- sqrt(Y)        # also square-root
ln.y <- log(Y)           # natural log
cubert.y <- Y^(1/3)      # cube-root
inv.y <- Y^(-1)          # inverse
inv2.y <- 1/Y            # also inverse
asin.y <- asin(sqrt(Y))  # arcsine square-root
```

[22] Box and Cox (1964) provided a statistical and graphical method for identifying the appropriate power transformation for the response variable. The details of this method are beyond the scope of this class but, in general, the method searches for a λ that minimizes the RSS (or SS_{within}). A slightly modified Box and Cox approach is implemented in R by sending a `lm` object to `boxcox()` from the `MASS` package.

2.6.3 Interpretations After Transformations

Care must be taken with interpretations following transformations. A few simple rules help in this regard:

1. Make sure to tell the reader what transformation you used and how you arrived at it.
2. Make sure that you refer to the transformed response variable in your conclusions (i.e., say "the mean square root of the response variable differed among treatment groups" rather than "the mean of the response variable").
3. The values should be back-transformed to the original scale when referring to means or confidence intervals of means.[23]

Back-Transformation Issues

Back-transformation is the process of reversing the results found on the transformed scale to the original scale for ease of interpretation. For example, log transformations are reversed with the exponential function and square root transformations are reversed by squaring the results. Wherever possible, back-transformations should be performed in order to provide results on the original scale of measurement. However, back-transformation must be considered carefully because the back-transformed result may be subject to systematic bias.

It is commonly known that back-transforming the mean value on the log scale underestimates the mean value on the original scale. This observation stems from the fact that the back-transformed mean value from the log scale is equal to the geometric mean[24] of the values on the original scale (Appendix **??**). The geometric mean is always less than the arithmetic mean[25] and, thus, the back-transformed mean always underestimates the arithmetic mean from the original scale. A wide variety of "corrections" for this back-transformation bias with logarithms have been suggested in the literature. The most common correction, derived from the analysis of normal and log-normal distributional theory, is to multiply the back-transformed value by

$$e^{\frac{MS_{Within}}{2}}$$

Another issue arises when back-transforming differences in means. For example, because the difference in the log of two values is equal to the log of the ratio of the two values, the back-transformed difference in two values becomes the ratio of the two values; i.e.,

$$e^{log(x_1)-log(x_2)} = e^{log(\frac{x_1}{x_2})} = \frac{x_1}{x_2}$$

Thus, a confidence interval for the difference in two means of a log-transformed variable becomes a confidence interval for the RATIO of two means on the original scale.

[23] This rule refers more to simple linear regression models where confidence intervals for slopes, which are not means, are of interest.

[24] The geometric mean is defined as the nth root of the product of the n values.

[25] The arithmetic mean is the sum of all values divided by n.

2.7 Example Analyses III

2.7.1 Peak Discharge

Mathematical models are used to predict flood flow frequency and estimates of peak discharge for the Mississippi River watershed. These models are important for forecasting potential dangers to the public. A civil engineer is interested in determining whether four different methods for estimating flood flow frequency produce equivalent estimates of peak discharge when applied to the same watershed. The statistical hypotheses to be examined are

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_A : \text{At least one pair of means is different}$$

where the subscripts identify the four different methods.

Data Collection

Each estimation method was used six times on the watershed and the resulting discharge estimates (in cubic feet per second) were recorded.

EDA & Assumption Checking

From the information given, the data do not appear to be independent either within treatments or among treatments (i.e., they are all on the same watershed). This assumption appears to be violated. However, the single watershed is the "population" of interest to the engineer. Thus, this form of data collection is not problematic unless the engineer (or you) attempt to make strict inferences to other watersheds. I will proceed with the analysis because of this last statement. The variances among treatments appear to be non-constant (Levene's $p = 0.0136$). The residual plot from the initial ANOVA fit also indicates a heteroscedasticity (Figure 2.16). This suggests the need for a transformation.

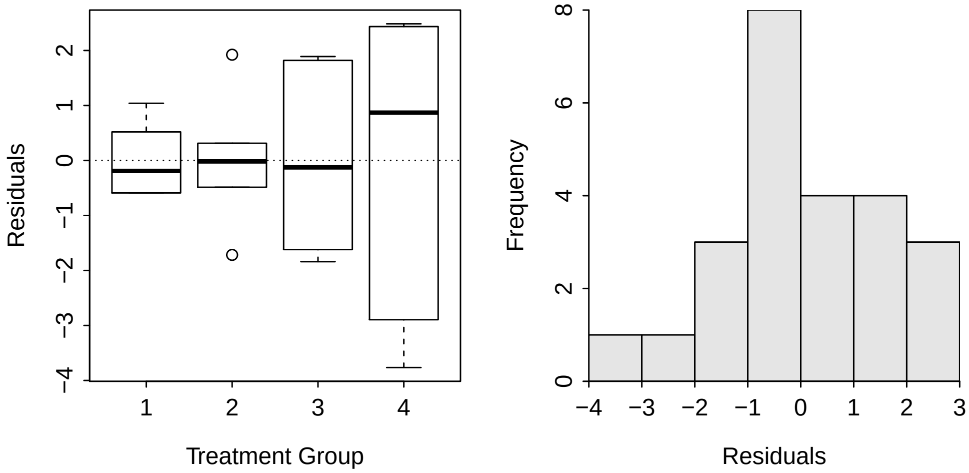

Figure 2.16. Residual plot (Left) and histogram of residuals (Right) from the one-way ANOVA on raw peak discharge data.

The data were examined with `transChooser()` and it was determined that a square root transformation may be appropriate. In fact, the square root transformation appeared to have stabilized the variances

(Levene's $p = 0.8680$; Figure 2.17) and normalized the residuals (Anderson Darling $p = 0.4207$). The largest Studentized residual had a very large p-value (>1) indicating that no significant outliers were present in these data. Thus, the assumptions of the ANOVA model appeared to have been adequately met on the square-root scale. Thus, the square root transformed data were examined with a one-way ANOVA.

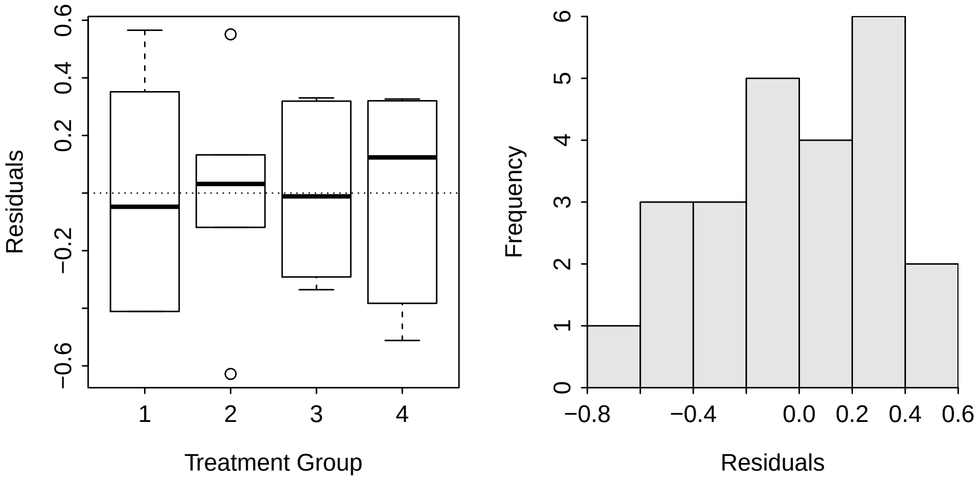

Figure 2.17. Residual plot (Left) and histogram of residuals (Right) from one-way ANOVA on square root transformed peak discharge data.

Results

There appeared to be a significant difference in mean square root peak discharge among the four methods ($p < 0.00005$; Table 2.8). Tukey's HSD multiple comparison method indicated that no two means were equal (Table 2.9) and, thus, the mean square-root of peak discharge increased significantly at each step from method 1 to method 4 (Figure 2.18).

Table 2.8. ANOVA results of square-root peak discharge for four methods.

```
          Df Sum Sq Mean Sq F value    Pr(>F)
method     3 32.684 10.8947  81.049 2.296e-11
Residuals 20  2.688  0.1344
```

Table 2.9. Tukey adjusted p-values for pairwise comparisons of square-root peak discharge for four methods.

```
           Estimate Std. Error   t value      p value
2 - 1 = 0 0.8245678  0.2116771  3.895403 4.568880e-03
3 - 1 = 0 2.0458777  0.2116771  9.665085 2.313437e-08
4 - 1 = 0 3.0634165  0.2116771 14.472118 6.328271e-15
3 - 2 = 0 1.2213099  0.2116771  5.769682 5.939964e-05
4 - 2 = 0 2.2388488  0.2116771 10.576715 7.395198e-10
4 - 3 = 0 1.0175389  0.2116771  4.807032 5.870941e-04
```

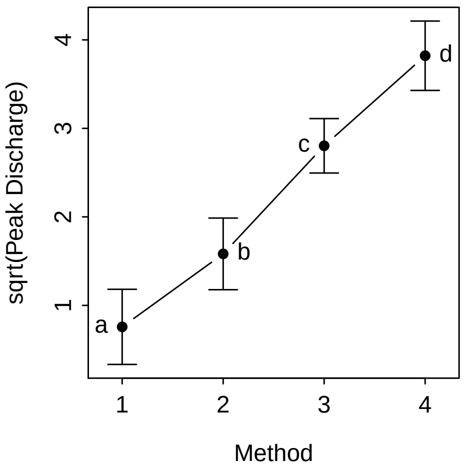

Figure 2.18. Plot of the mean square root transformed peak discharge data with 95% confidence intervals and significance notations.

Conclusion

The four methods produced significantly different mean square-root peak discharge estimates. The methods when ranked from lowest to highest estimates are as follows: method 1, method 2, method 3, and method 4. Broader inferences cannot be made because there appears to be no randomization in this experimental design (at least from the information that is given).

Appendix – R commands

```
PeakDischarge <- read.csv("PeakDischarge.csv")
PeakDischarge$method <- factor(PeakDischarge$method)
pd.lm <- lm(discharge~method,data=PeakDischarge)
levenesTest(pd.lm)
residPlot(pd.lm)
transChooser(pd.lm)
PeakDischarge$sqrtdis <- sqrt(PeakDischarge$discharge)
pd.lm1 <- lm(sqrtdis~method,data=PeakDischarge)
levenesTest(pd.lm1)
adTest(pd.lm1$residuals)
residPlot(pd.lm1)
outlierTest(pd.lm1)
anova(pd.lm1)
pd.mc1 <- glht(pd.lm1,mcp(method="Tukey"))
summary(pd.mc1)
fitPlot(pd.lm1,xlab="Method",ylab="sqrt(Peak Discharge)")
addSigLetters(pd.lm1,c("a","b","c","d"),pos=c(2,4,2,4))
```

2.8 Summary Process

The process of fitting and interpreting linear models is as much an art as it is a science. The "feel" for fitting these models comes with experience. The following is a process to consider for fitting a one-way ANOVA model. Consider this process as you learn to fit one-way ANOVA models, but don't consider this to be a concrete process for all models.

1. Perform a thorough EDA of the quantitative response variable.
   - Pay close attention to the distributional shape, center, dispersion, and outliers within each level of the factor variable.
2. Fit the untransformed ultimate full model [`lm()`].
3. Check the assumptions of the fit of the untransformed model.
   - Check equality of variances with a Levene's test [`levenesTest()`] and residual plot [`residPlot()`].
   - Check normality of residuals with an Anderson-Darling test [`adTest()`] and histogram of residuals [`residPlot()`].
   - Check for outliers with an outlier test [`outlierTest()`], residual plot, and histogram of residuals.
4. If an assumption or assumptions are violated, then attempt to find a transformation where the assumptions are met.
   - Use the trial-and-error method [`transChooser()`], theory, or experience to identify a possible transformation.
   - If only an outlier exists (i.e., equal variances and normal residuals) and no transformation corrects the "problem" then consider removing the outlier from the data set.
   - Fit the ultimate full model with the transformed response or reduced data set.
5. Construct an ANOVA table for the full model [`anova()`] and interpret the overall F-test.
6. If differences among level means exist, then use a multiple comparison technique [`glht()`] to identify specific differences.
   - Use Tukey's HSD method if comparing all possible pairs of means.
   - Use Dunnett's method if comparing all group means to one specific group mean (e.g., a control).
7. Summarize findings with significance letters [`addSigLetters()`] on a means plot [`fitPlot()`] or table [`Summarize()`].