Readings for MTH207

Derek H. Ogle

2020-12-22

Contents

Preface

This BOOK-IN-PROGRESS was last updated on: 22 Dec 2020.

FOUNDATIONS

Chapter 1

Model Types & Methods

During this course we will examine a variety of models called either *general linear* or *generalIZED linear* models. General linear models have a quantitative response variable and generally assume that the "errors" around the model follow a normal distribution. General linear models that we will discuss are a **One-Way ANOVA**[1], **Two-WAY ANOVA**, **Simple Linear Regression**, and **Indicator Variable Regression**. GeneralIZED linear models do not require a quantitative response variable nor "errors" that are normally distributed. Thus, generalIZED linear models are more flexible than general linear models. The only generalIZED linear model that we will encounter in this course is **Logistic Regression**, but the chi-square test from your introductory statistics course can also be cast as a generalIZED linear model.

Response Variable: The variable thought to depend upon, be explained by, or be predicted by other variables.

All models covered in this course will have **only one** response variable

Both general and generalIZED linear models can have a single explanatory variable that can be either quantitative or categorical, or multiple explanatory variables that can be all quantitative, all categorical, or a mixture of both quantitative and categorical. Ultimately, there can be several explanatory variables in a model, but we will only consider one or two explanatory variables in this course.

Explanatory Variable: A variable thought to explained or be able to predict the response variable.

[1] ANOVA is short for ANalysis Of VAriance

Table 1.1: Response and explanatory variables types for the linear models considered in this course.

Response	Explanatory	Linear Model
Quantitative	Categorical (only one)	One-Way ANOVA
Quantitative	Categorical (two)	Two-Way ANOVA
Quantitative	Quantitative (only one)	Simple Linear Regression
Quantitative	Quantitative (one) & Categorical (one)	Indicator Variable Regression
Binomial	Quantitative (or Both)	(Binary) Logistic Regression

1.1 Distinguishing Methods

The five methods that will be covered in this course can be distinguished by considering only the type of response variable and the types and number of explanatory variables (Table 1.1). Thus, you will want to review variable types and definitions and distinctions of response and explanatory variables from your introductory statistics course.

1.2 Method Purposes

As seen above, each method uses different types of data. Not surprisingly then, each method is used to test different hypotheses or has a different analytical purpose. These purposes will be discussed in detail in subsequent modules. However, the major objective of each method is explained briefly below (in the order that we will cover them).

Each example uses a data set that contains data about mirex concentrations (`mirex`) for two species of salmon (`species`) captured in six years between 1977 and 1999 (`year`) in Lake Ontario. The weight of each fish (`weight`) and whether or not the mirex concentration exceeded the EPA limit of 0.1 mg/kg (`exceeds_limit`) were also recorded.

A **one-way ANOVA** is used to determine if the means of the quantitative response variable (`mirex`) differ among two or more groups defined by a single categorical variable (e.g., `year`).

A **two-way ANOVA** is used to determine if the means of the quantitative

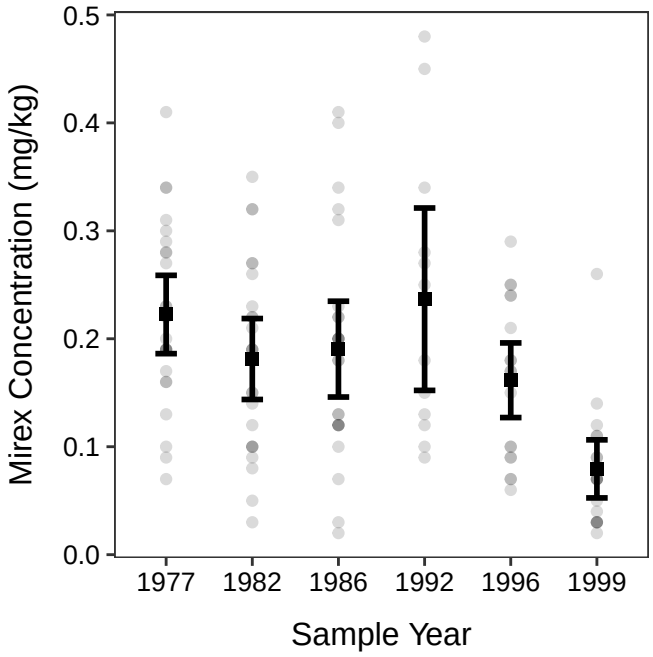


Figure 1.1: Mean mirex concentration by sample year. This is an example of a One-Way ANOVA.

response variable (`mirex`) differ among groups of one categorical variable (e.g., `year`), among groups of another categorical variable (e.g., `species`), or by the interaction between the two categorical variables.

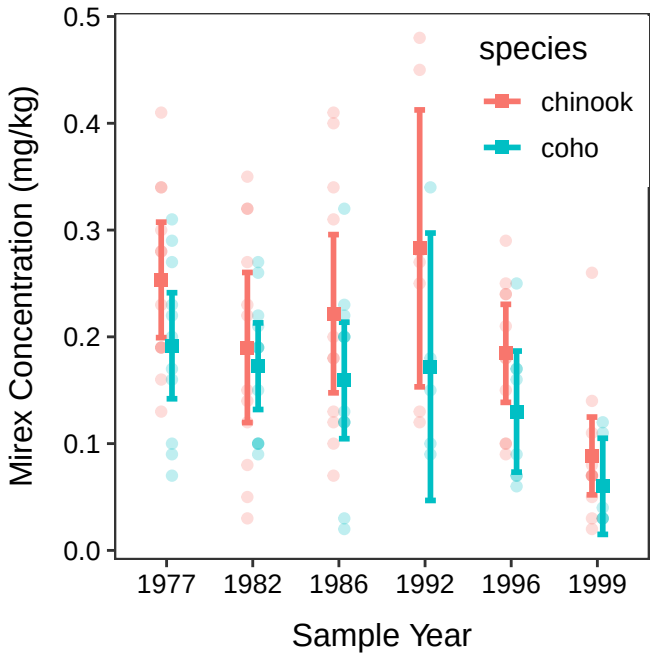


Figure 1.2: Mean mirex concentration by sample year and salmon species. This is an example of a Two-Way ANOVA.

A **simple linear regression** is used to determine if there is a relationship between the quantitative response variable (e.g., `mirex`) and a single quantitative explanatory variable (e.g., `weight`).

An **indicator variable regression** is used to determine if the relationship between a quantitative response (e.g., `mirex`) and a quantitative explanatory variable (e.g., `weight`) differs between two or more groups defined by a categorical explanatory variable (e.g., `species`). This will look like two (or more) simple linear regressions are being compared.

A **logistic regression** is used to determine if there is a relationship between the probability of "success" for a binary[2] categorical response variable (e.g.,

[2]Binary means there are only two categories – generically "success" and "failure".

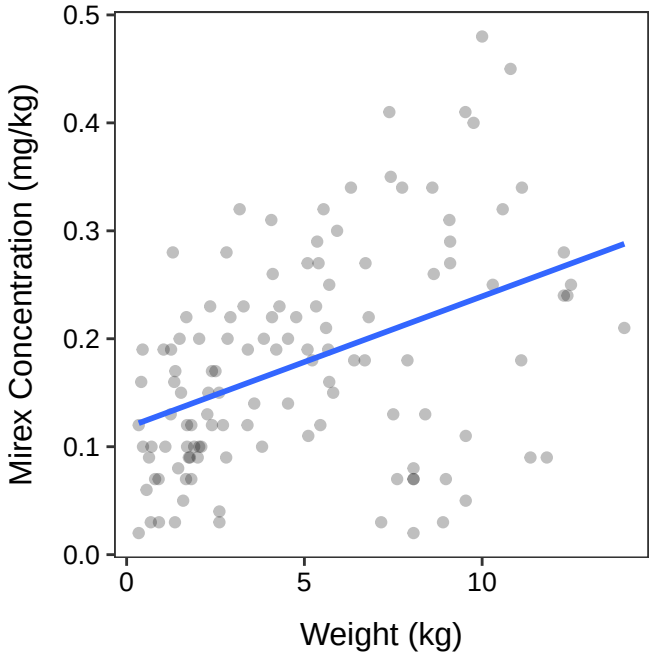


Figure 1.3: Mirex concentration by fish weight. This is an example of a Simple Linear Regression.

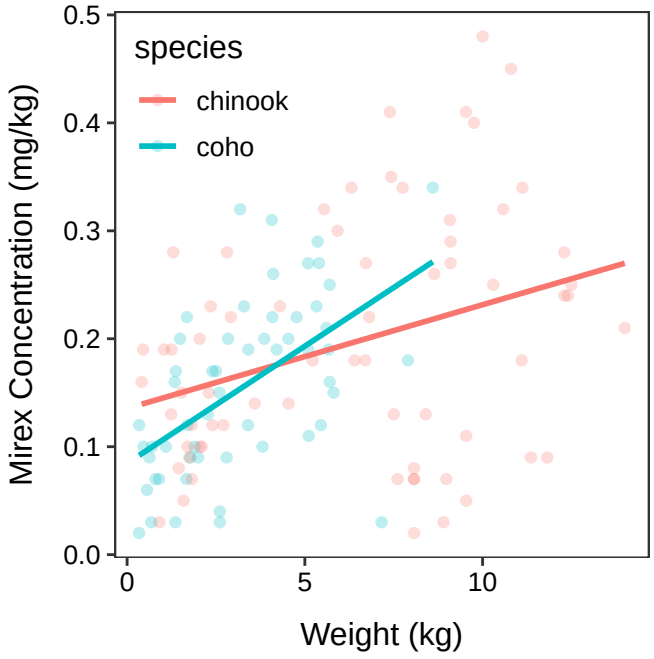


Figure 1.4: Mirex concentration by fish weight seprated by salmon species. This is an example of an Indicator Variable Regression.

`exceeds_limit`) and the quantitative explanatory variable (e.g., `weight`).

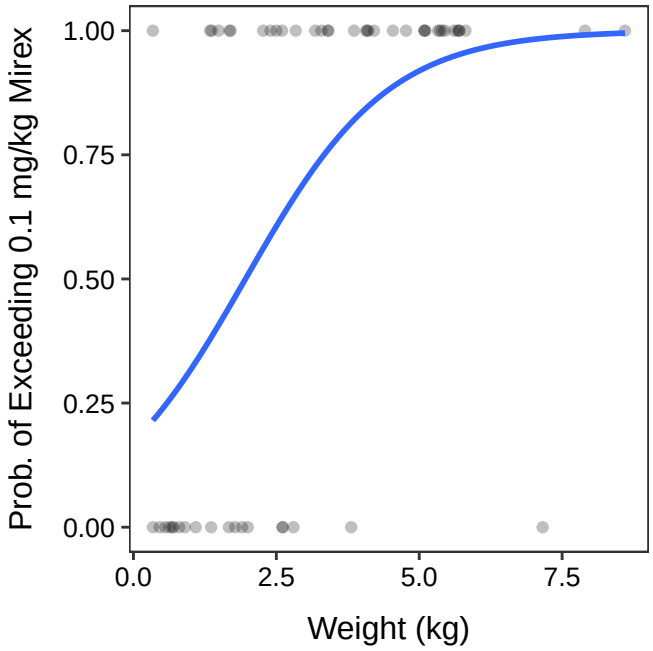


Figure 1.5: The probability that the mirex concentration exceed the 0.1 mg/kg threshold by fish weight. This is an example of a Logistic Regression.

From these examples it should be apparent that "ANOVAs" are about **comparing means** among groups and will look like means (usually with confidence intervals) plotted as points for each group. In contrast "regressions" are about exploring **relationships** and will look like a line or a curve when plotted.

ANOVAs compare means; regressions examine relationships.

Chapter 2

2-Sample t Review

A two-sample t-test is a statistical method for comparing the means of a quantitative variable between two populations represented by two independent samples. The specific details of a two-sample t-test were covered in your introductory statistics course and will only be cursorily reviewed here.

2.1 Review

The null hypothesis for a 2-sample t-test is H0: $\mu_1 = \mu_2$, where μ is the population mean and the subscripts represent the two populations. The alternative hypothesis of a 2-sample t-test may be "less than", "greater than", or "not equals". We will use HA: $\mu_1 \neq \mu_2$ for most examples in this course.

The 2-sample t-test assumes that (i) individuals in the populations are independent; (ii) the sample size (n) is great than 40, greater than 15 and the histograms are not strongly skewed, or the histograms are normally distributed; and (iii) the population variances are equal. The assumption of equal variances for the 2-sample t-test is tested with Levene's test, which uses H0: $\sigma_1^2 = \sigma_2^2$ and HA: $\sigma_1^2 \neq \sigma_2^2$, where σ^2 is the population variance. If H0 is rejected for Levene's test then the variances for both populations are assumed to be equal, such that only one combined sample variance needs to be estimated. That combined sample variance is called the *the pooled sample variance* and is computed as a weighted mean of the two sample variances,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

If the three assumptions are met then the statistic for the 2-sample t-test is $\bar{x}_1 - \bar{x}_2$ which is immediately standardized to a t test statistic with

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

The t test statistic is converted to a p-value using a t-distribution with n_1+n_2-2 df. Of course, a p-value $<$ means that H0 is rejected and the two population means appear to be different. A confidence interval would then be used to fully describe which population mean was greater (or smaller) and by how much.

2.2 Analysis in R

2.2.1 Data Format

Data for a 2-sample t-test must be in stacked format, where measurements are in one column and a label for the populations is in another column. Each row corresponds to the measurement and population of a single individual.

The data (data, meta) for the example below are the biological oxygen demands (BOD) at the inlet and outlet to an aquaculture facility. These data illustrate stacked data because each row is one water sample with two variables recorded – BOD and where the sample came from.

```
aqua <- read.csv("BOD.csv")
```

```
headtail(aqua)
```

```
#R>        BOD    src
#R>   1  6.782  inlet
#R>   2  5.809  inlet
#R>   3  6.849  inlet
#R>   18 8.545 outlet
#R>   19 8.063 outlet
#R>   20 8.001 outlet
```

Stacked Data: Data where the quantitative measurements of two or more groups are "stacked" on top of each other and a second variable is used to record to which group (or population) the measurement belongs.

Stacked data is required for the methods used in this course.

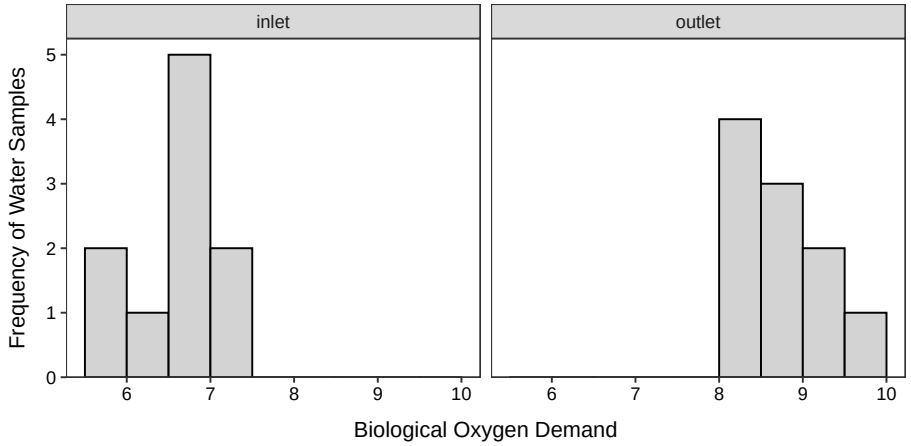
Specific details for performing a 2-sample t-test in R were provided in your introductory statistics course, but will be cursorily reviewed below.

2.2.2 Assumption Checking

The metadata suggests that measurements at the intake and outtake were taken at different times. Thus, there is no reasonable reason to think that individuals are dependent across the two populations. Thus, the independence assumption is met.

The sample size is less than 40 but greater than 15. The histograms shown below are not particularly informative because of the small sample size. The histogram for the inlet samples appears to be not strongly skewed, but that for the outlet appears to be strongly right-skewed. I am going to continue with this analysis, but I will be cautious with my final interpretations.

```
ggplot(data=aqua,mapping=aes(x=BOD)) +
  geom_histogram(binwidth=0.5,boundary=0,color="black",fill="lightgray") +
  labs(y="Frequency of Water Samples",x="Biological Oxygen Demand") +
  scale_y_continuous(expand=expansion(mult=c(0,0.05))) +
  theme_NCStats() +
  facet_wrap(vars(src))
```



The `ggplot2` package is required to make plots with `ggplot()`.

Levene's test is computed with `levenesTest()` using a formula of `response~groups` as the first argument, where `response` represents the name of the quantitative response variable and `groups` represents the name of the categorical variable that identifies the two populations. The data.frame with the variables must be in `data=`. From the results below, it is concluded that the population variances appear to be equal because the Levene's test p-value (0.5913) is greater than $=0.05$.

```
levenesTest(BOD~src,data=aqua)
```

```
#R>   Levene's Test for Homogeneity of Variance (center = median)
#R>         Df F value Pr(>F)
#R>   group  1  0.2989 0.5913
#R>         18
```

Levene's test requires the `NCStats` package to be loaded.

2.2.3 Analysis

A 2-sample t-test is constructed in R with `t.test()` using the exact same `response~groups` formula and `data=` used in `levenesTest()`. Additionally, `var.equal=TRUE` is used when the two population variances should be considered equal. By default `t.test()` uses a "not equals" HA and a 95% confidence interval. In the results below the two sample means are 6.6538 for the inlet group and 8.6873 for the outlet group such that the statistic is 6.6538-8.6873=-2.0335; the t test statistic is -8.994 with 18 df; and the p-value is <0.00005 (or, more specifically, 4.449e-08).[1] Because the p-value< the H0 is rejected and we conclude that the mean BOD at the inlet is lower than the mean BOD at the outlet. More specifically, the mean BOD at the inlet is between 1.558 and 2.509 units lower than the mean BOD at the outlet. Thus, it appears that the mean BOD in the water is increased from when it enters to when it leaves the aquaculture facility.

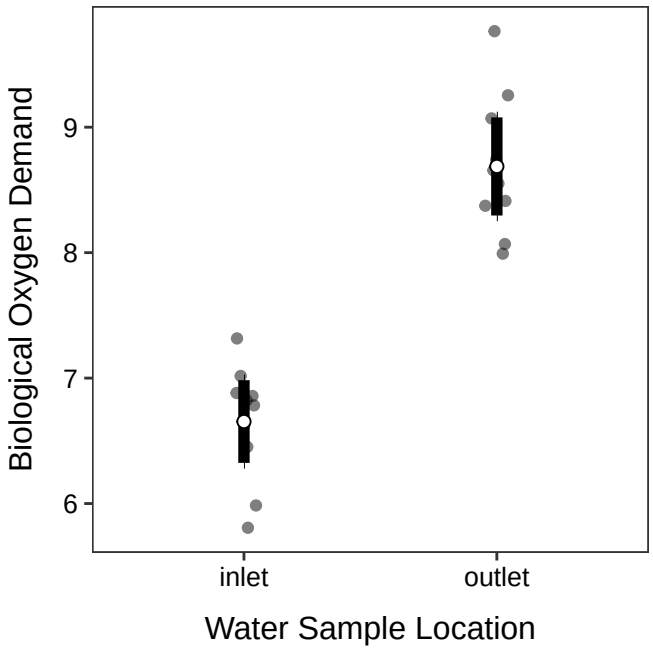
```
t.test(BOD~src,data=aqua,var.equal=TRUE)
```

```
#R>    Two Sample t-test with BOD by src
#R>   t = -8.994, df = 18, p-value = 4.449e-08
#R>   alternative hypothesis: true difference in means is not equal to 0
#R>   95 percent confidence interval:
#R>    -2.508511 -1.558489
#R>   sample estimates:
#R>    mean in group inlet mean in group outlet
#R>                 6.6538               8.6873
```

A graphic that illustrates the mean BOD with 95% confidence intervals for each sampling location is constructed below. Note that in the code below that the only items you need to change for your own data is in the first line, where `data=` should be set to the name of your data and `x=` and `y=` should be set to the names of the explanatory and response variables, respectively.

[1] I usually round my p-values to four decimal places. In this case that would mean 0.0000 which is awkward. Thus, I will say p<0.00005 as the fifth position must have been less than 5 to round to 0.0000.

```
ggplot(data=aqua,mapping=aes(x=src,y=BOD)) +
  geom_jitter(alpha=0.5,width=0.05) +
  stat_summary(fun.data=mean_cl_normal,geom="errorbar",size=2,width=0) +
  stat_summary(fun=mean,geom="point",pch=21,fill="white",size=2) +
  labs(x="Water Sample Location",y="Biological Oxygen Demand") +
  theme_NCStats()
```



2.3 Signal-to-Noise

The ratio of signal to noise can be a useful metaphor for understanding hypothesis testing, as we have done here, and model comparisons, as we will do in future modules. In this metaphor, think of "signal" as how different two things are and "noise" as anything that gets in the way of you receiving the signal. For example, the difference in heights of two students standing on the other side of the room is a "signal", but smoke in the room that makes it difficult to see those students is "noise." As another example, it may be easy to see an orange kayak (the "signal") on Lake Superior on a calm day but harder to see it on a wavy day (i.e., more "noise").

In a 2-sample t-test, the "signal" is the difference in the two group means (Figure 2.1), which is measured by $\bar{x}_1 - \bar{x}_2$, the numerator of the t-test statistic.

The bigger the difference in sample means the stronger the "signal" that the population means are different.

The "signal" is the difference in sample means

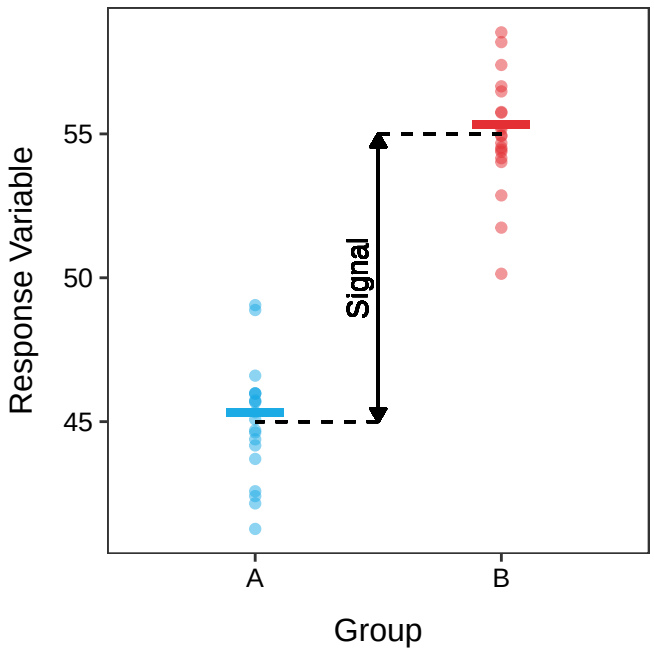


Figure 2.1: Response variable by group for each indiviual (points) with group means shown as horizontal segments. The difference in sample means is highlighted as the "signal" in these data.

"Noise" is sampling variability, the fact that statistics (e.g., $\bar{x}_1$ and $\bar{x}_2$) vary from sample to sample. Sampling variability in a 2-sample t-test is measured by $SE_{\bar{x}_1-\bar{x}_2}$, which is the denominator of the t test statistic, or $\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$. This SE increases with increasing s_p^2 and decreases with increasing n1 and n2. So the "noise" increases as the natural variability of individuals around their group means (i.e., s_p^2) increases (Figure 2.2), but decreases as the sample size increases.

The "noise" is sampling variability

The ratio of signal to noise is related to whether we will be able to detect the difference between two things or not. If the signal is large relative to the noise then the signal will be detected. In other words, will be able to tell the difference

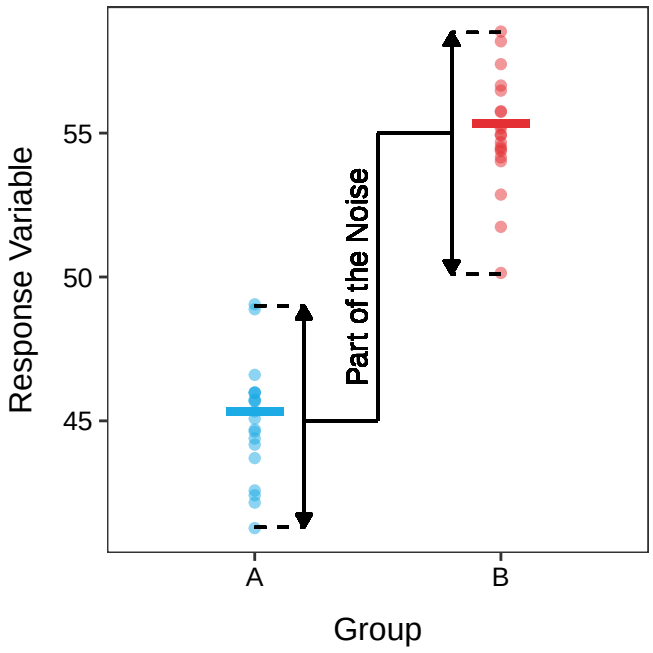


Figure 2.2: Response variable by group for each indiviual (points) with group means shown as horizontal segments. The variability of individuals around the group means is highlighted as a part of the "noise" in these data.

in heights of students if the room is not full of smoke.

For example, each panel in Figure 2.3 has the same signal (difference in means) but the noise (i.e., SE) increases from left to right. In the left-most panel it is very clear that the sample means are different (high signal-to-noise ratio), but in the right-most panel it is less clear that the sample means are different (low signal-to-noise ratio).

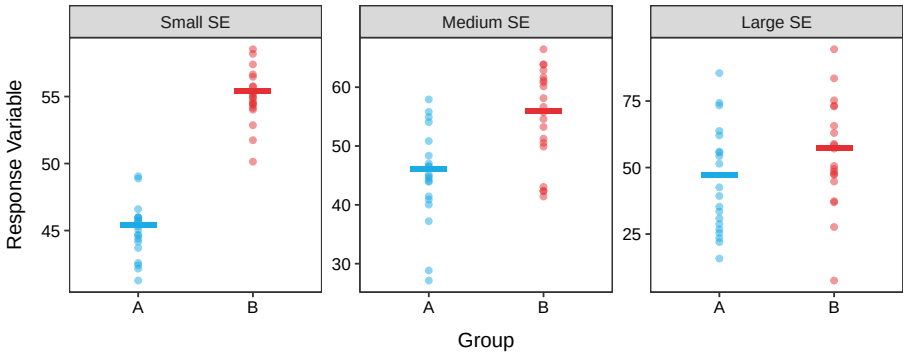


Figure 2.3: Response variable by group for each indiviual (points) with group means shown as horizontal segments for three different standard errors (SE; i.e., "noise"). Note that the group means are the same in all three panels.

The t test statistic is a measure of signal (i.e., difference in sample means) to noise (i.e., sampling variability as measured by the SE)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\text{Signal}}{\text{Noise}}$$

Thus, larger values of the t test statistic indicate a larger signal-to-noise ratio. Larger t test statistics are further into the tail of the t distribution and result in smaller p-values. Therefore, small p-values represent larger signal-to-noise ratios and are more likely to lead to concluding that the population means differ. In other words, you were able to detect the "signal" through the "noise."

More signal-to-noise means smaller p-values

We will return to the signal-to-noise metaphor throughout this course.

Chapter 3

Model Concepts

3.1 What is a Model

A model is a representation of something or some phenomena. It is usually a simplification or an abstraction that helps our understanding of the more complex reality. A mathematical or statistical model is an equation or system of equations that is meant to characterize the general characteristics of observations. Statistical models do not represent every observation perfectly, rather they attempt to best represent the "central tendency" of the observations. Weather forecasts are based on mathematical and statistical models. You have observed at least two statistical models in your introductory statistics course – the mean and the regression line (Figure 3.1).

Models can predict an observation but generally not perfectly. For example, weather forecasters predict the temperature for tomorrow but will most likely be off by (hopefully only) a small amount.

An observed value of the response variable can be thought of as being equal to a value predicted from a model plus some deviation, or error, from that prediction; i.e.,

$$\text{Observed Response} = \text{Model Predicted Response} + \text{error}$$

For example, tomorrow's temperature may be 74oF, which is the predicted 76oF from the forecaster's model plus -2oF "error."

In statistics, one model for predicting the response variable for an individual in a group is to use the mean for the group. My best guess at the height of an unknown student is to guess that they are average for "their group." Obviously, most individuals are not truly average, so the specific individual will deviate

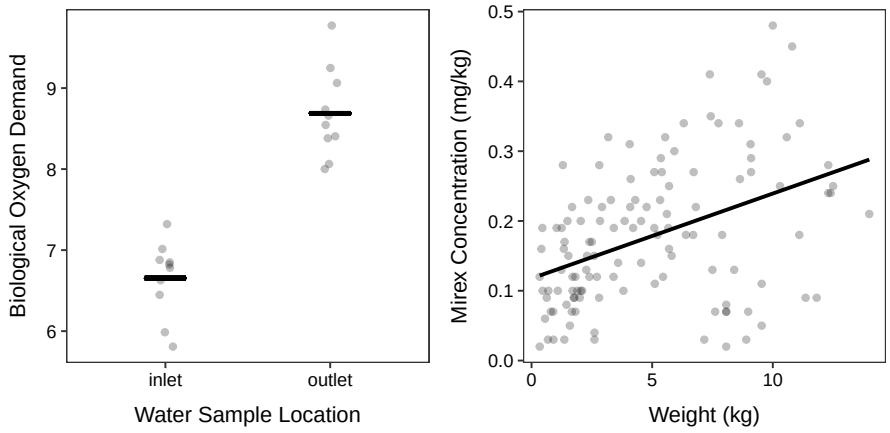


Figure 3.1: Two examples of models seen in your introductory statistics course – two means (Left) and regression line (Right).

from the mean. In Figure 3.2 an observation is shown as a red point, the predicted value for that individual is shown as a horizontal line at the mean for the individual's group, and the "error" from this prediction is shown as the vertical red line.

We always predict the **response** variable with a model.

In the context of a simple linear regression, the predicted value is obtained by plugging the observed value of the explanatory variable into the regression equation. Thus, the "error" is the vertical distance between an observed point and the corresponding point on the line (Figure 3.3).

Many hypothesis tests, including the two-sample t-test, can be cast in a framework of competing statistical models. Using this framework requires assessing the relative fit (to data) and complexity of a model. The remainder of this module is about measuring fit and complexity of models. We will discuss fit and formally compare two models to see which is "best" in the next module.

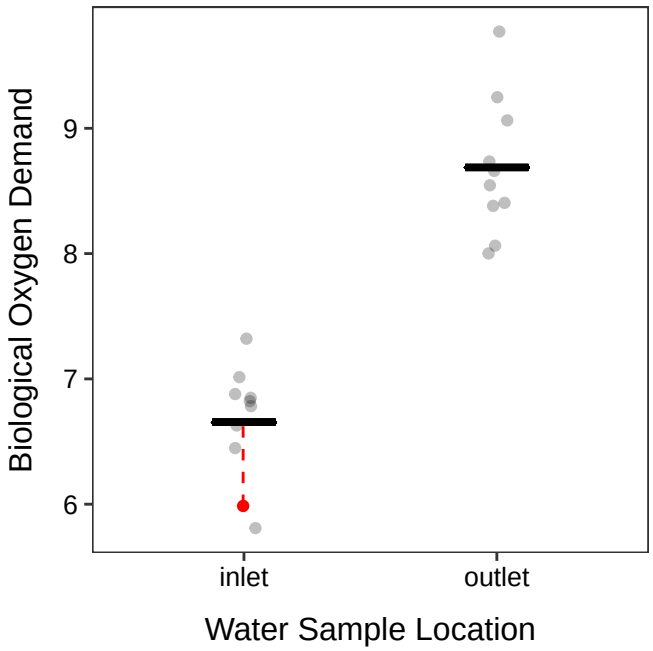


Figure 3.2: Biological oxygen demand versus sample location (points) with group means shown by horizontal segments. The residual from a model that uses a separate mean for both groups is shown.

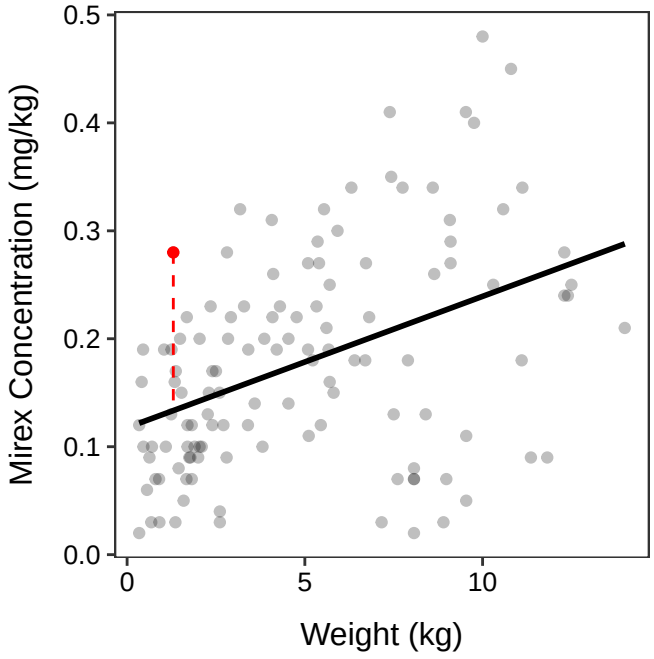


Figure 3.3: Mirex concentration versus fish weight with a simple linear regression line show. The residual from the regression line model is shown.

3.2 Assessing Fit (SS)

3.2.1 A Residual

A residual is an estimate of the "error" discussed in the previous section. If you rearrange the formula shown above and replace "error" with "residual" you see that

$$\text{residual} = \text{Observed Response} - \text{Model Predicted Response}$$

Visually a residual is the vertical distance between a point and the "model", as shown by the vertical dashed lines above (or further below). Residuals are vertical distances because they are the difference between two values of the response variable, which is always plotted on the y-axis.

Residuals are *vertical* distances between an observation and the model.

Residuals are negative if the point is "below" the model prediction and positive if the point is "above" the model prediction. More importantly, the absolute value of the residual is a measure of how close the model prediction is to the point or how well the model fits the individual point. Large residuals (in an absolute value sense) mean that the point is far from the model prediction and, thus, the model does not represent that point very well. Points with small residuals, in contrast, are near the model prediction and are thus well-represented by the model. Figure 3.4 shows points with relatively large residuals in red and relatively small residuals in blue.

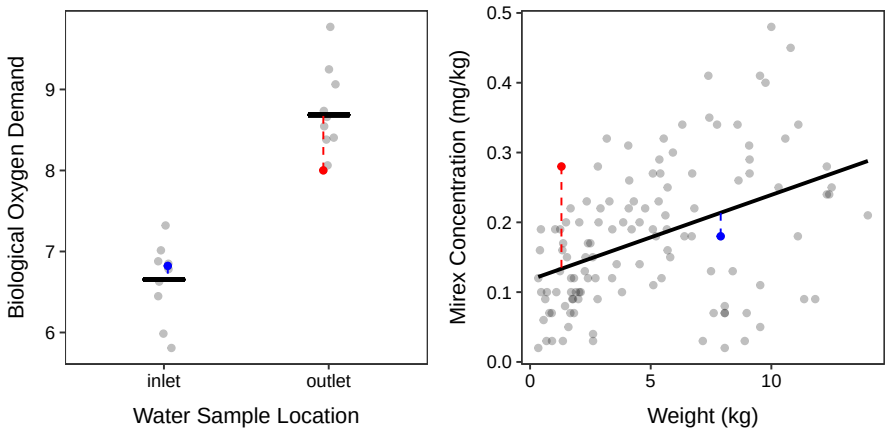


Figure 3.4: Same plots as previously but with a "large" residual shown in red and a "small" residual shown in blue.

3.2.2 Residual Sum-of-Squares

If a residual measures how closely a model comes to a point then it stands to reason that the sum of all of the residuals measures how closely a model comes to all of the points. Unfortunately, because residuals are negative and positive they always sum to 0.[1] Thus, the sum of all residuals is not a useful measure of the overall fit of a model.

Instead of summing residuals, statisticians sum squared residuals into a quantity called a **residual sum-of-squares (RSS)**.[2] Using the formula for a residual from above, an RSS for a given set of observed data and a model is computed with

$$\text{RSS} = \sum_{data} (\text{Observed Response} - \text{Model Predicted Response})^2$$

The RSS measures how closely the model comes to *all* of the observations.

The RSS is on an unfamiliar scale (squared residuals?) but it maintains the same conceptual idea that summing residuals would have. Mainly, the smaller the RSS the more closely the points are to the model. The full set of residuals required to compute an RSS are shown in Figure 3.5.

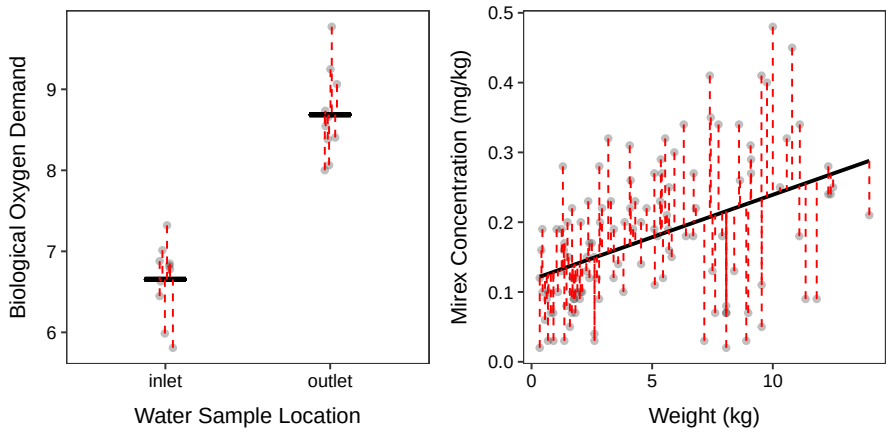


Figure 3.5: Same plots as previously but with all residuals shown.

As a value, the RSS is a measure of how *poorly* the model fits the data – i.e., small values are a good fit, large values are a poor fit. Thus, the RSS is often

[1] Under certain reasonable assumptions.

[2] Some statisticans call this an error sum-of-squares or a sum of squared errors (SSE)

called "a measure of lack-of-fit" of the model to the observations.

An RSS is a measure of the "lack-of-fit" of a model to the data.

Unfortunately, the magnitude of the RSS is only useful in comparison to other RSS computed for different models from the same data. We will discuss this further in the next module.

3.3 Residual Degrees-of-Freedom

You used degrees-of-freedom (df) with t-tests and chi-square tests in your introductory statistics course. However, you likely did not discuss what degrees-of-freedom mean and where they come from. I will discuss this briefly here, but we will use df more in the next module.

Residual degrees-of-freedom (Rdf) are the number of observations that are "free" to vary if the sample size (n) and number of parameters estimated is known. As a simple example, suppose that we know that $\bar{x}$=13 from n=4 observations. With just this information can I tell you the values for the four observations that went into $\bar{x}$? Clearly I cannot. If you give me one observation can I tell you the remaining three? No! If you tell me two? No! If you tell me three observations can I tell you the last observation? Yes, because the total of the four numbers must be 52 (=$n\bar{x}$=4×13); so the last number must be 52 minus the total of the three numbers you told me. In this case, three numbers were "free" to be any value before the last number was set. Thus, this case has three residual degrees-of-freedom.

Residual degrees-of-freedom are more complicated to explain in other situations, but generally

$$\text{Rdf} = \text{Number of Observations} - \text{Number of Model Parameters}$$

In the example above, there were four observations (n) and one model parameter – $\bar{x}$ – so df=4-1=3. In Figure 3.1-Left there are 20 observations and two parameters (i.e., two group means) so Rdf=20-2=18. In Figure 3.1-Right there are 122 observations and two parameters (i.e., the slope and intercept of the regression line) so Rdf=122-2=120.

As a general rule, parameter estimates are more precisely estimated with more residual degrees-of-freedom. Thus, models that "preserve" residual degrees-of-freedom (i.e., have fewer parameters) are preferred, all else being equal.

3.4 Mean-Squares

Sums-of-squares are useful measures of model fit, but they are largely uninterpretible on their own. However, if a sum-of-squares is divided by its corresponding degrees-of-freedom it is called a **Mean-Square (MS)**. Mean-squares are the **variance** (i.e., squared standard deviation) of individuals around a given model. Mean-squares have useful mathematical properties as you will see in future modules. However, visually the square root of a mean square loosely describes how far each point is from the model (i.e., the "errors"), on average. The mean-squares are thus a measure of the "noise" around each model.

MS are variances; thus, the square root of MS are standard deviations

Chapter 4

Model Comparison

4.1 Competing Models

4.1.1 General

Many hypothesis tests can be cast in a framework of competing models. In this module we will cast the familiar 2-sample t-test in this framework which will then serve as a conceptual foundation for all other linear models in this course.

The two competing models are generically called the *simple* and *full* models (Table 4.1). The simple model is simpler than the full model in the sense that it has fewer parameters. However, the simple model fits the data "worse" than a full model. Thus, determining which model to use becomes a question of balancing "fit" (full model fits better than the simple model) with complexity (simple model is less complex than the full model). Because the simple model corresponds to H0 and the full model corresponds to HA, deciding which model to use is the same as deciding which hypothesis is supported by the data.

Table 4.1: Differences between the two generic model types.

Model	Parameters	Residual df	Relative Fit	Residual SS	Hypothesis
Simple	Fewer	More	Worse	Larger	Null
Full	More	Less	Better	Smaller	Alternative

4.1.2 2-Sample t-Test

Recall that H0 in a two-sample t-test is that the two population means do not differ (i.e., they are equal). In this hypothesis, if the two means do not differ than a single mean would adequately represent both groups. The general model from Section 3.1 [1] could be specified for this situation as

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

where Y_{ij} is the jth observation of the response variable in the ith group, μ is the population grand mean, and ϵ_{ij} is the "error" for the jth observation in the ith group. This model means that is the predicted response value for each observation and the model looks like the red line in Figure 4.1.

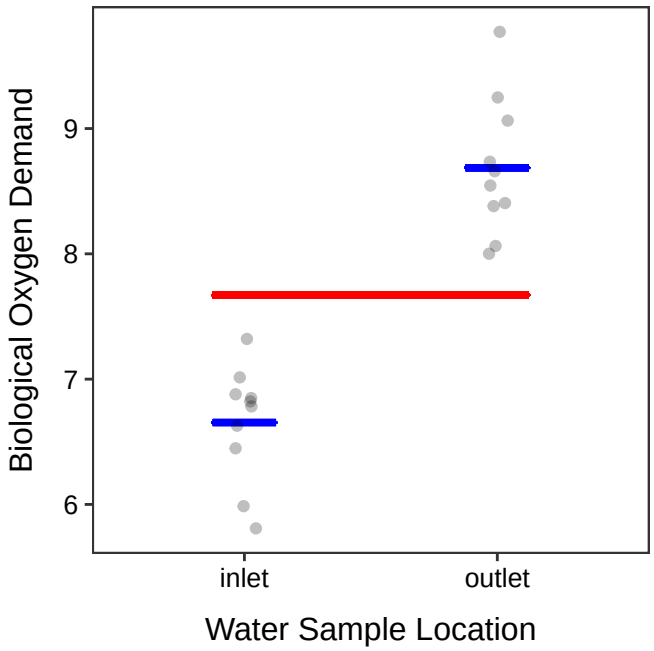


Figure 4.1: Biological oxygen demand versus sample location with group means shown as blue horizontal segments and the grand mean shown as a red horizontal segment.

In contrast, HA in the 2-sample t-test is that the two population means differ (i.e., they are not equal). This hypothesis suggests that two separate means are needed to predict observations in the separate groups. The model for this situation is

[1] Generally, Observation = Model Prediction + Error

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

where μ_i is the population mean for the ith group. This model means that μ_1 is the predicted response value for observations in the first group and μ_2 is the predicted response value for observations in the second group. This model looks like the two blue lines in the figure above.

Thus, for a 2-sample t-test, the **simple model** corresponds to H0: $\mu_1 = \mu_2$ ($=\mu$), has fewer parameters (i.e., requires only one mean; the red line in the plots above), and fits "worse."[2] In contrast, the **full model** corresponds to HA: $\mu_1 \neq \mu_2$, has more parameters (i.e., requires two means; the blue lines in the plots above), and fits "better."

In the ensuing sections we will develop a method to determine if the increase in "fit" is worth the increase in "complexity."

4.2 Measuring Increase in Fit

4.2.1 SSTotal and SSWithin

In Section 3.2.2 the idea of using the residual sum-of-squares (RSS) to measure the lack-of-fit of a model was introduced. Here we apply that concept to measure the lack-of-fit of the simple and full models, which we will then use to see how much "better" the full model fits than the simple model.

Remember that the full model will always fit better than the simple model, even if by just a small amount.

The RSS for the simple model using just the grand mean is called SSTotal and is computed with

$$\text{SS}_{\text{Total}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{\cdot\cdot}\right)^2$$

where I is the number of groups (=2 in a 2-sample t-test), n_i is the sample size in the ith group, and $\bar{Y}_{\cdot\cdot}$ is the sample grand mean as computed with

$$\bar{Y}_{\cdot\cdot} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij}}{n}$$

[2] We will be more objective in the following sections, but an examination of the plot above clearly shows that the red line does not represent the observations well.

where n is the sample size across all groups. The $\bar{Y}_{..}$ is used here because it is an estimate of the population grand mean, μ, which is used to make predictions in this simple model.

The formula for SSTotal may look daunting but it is just the sum of the squared residuals computed from each observation relative to the grand mean (Figure 4.2).

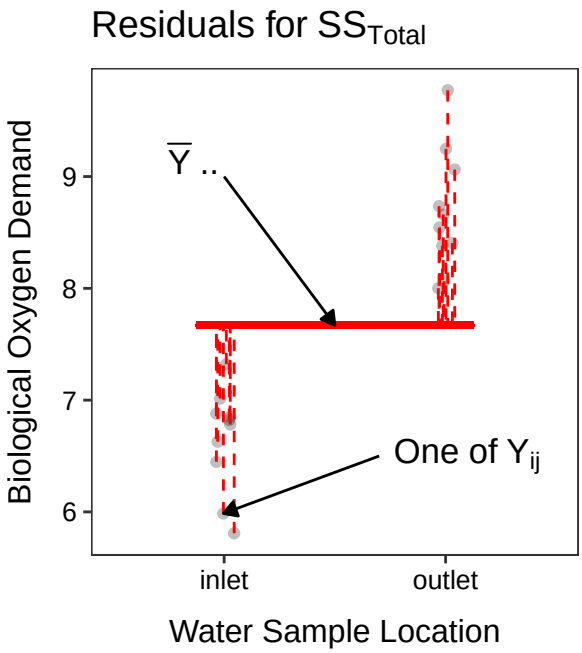


Figure 4.2: Biological oxygen demand versus sample location with the grand mean shown as a red horizontal segment. Residuals from the grand mean are shown by red vertical dashed lines. The sum of these residuals is SS_{Total}.

The RSS for the full model using separate group means is called SSWithin and is computed with

$$\mathrm{SS}_{\mathrm{Within}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$$

where $\bar{Y}_{i\cdot}$ are the sample group means as computed with

$$\bar{Y}_{\cdot\cdot} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$$

The $\bar{Y}_{i\cdot}$ are used here because they are an estimate of the population group means, μ_i, which are used to make predictions in this full model. Again, the formula for SSWithin may look imposing but it is just the sum of the squared residuals computed from each observation to the observation's group mean (Figure 4.3).

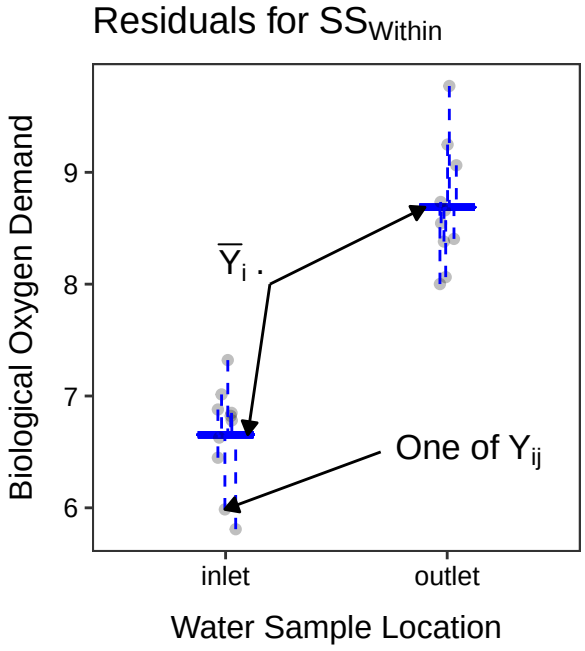


Figure 4.3: Biological oxygen demand versus sample location with the group means shown as blue horizontal segments. Residuals from the group means are shown by blue vertical dashed lines. The sum of these residuals is SS_{Within}.

Thus, SSTotal measures the lack-of-fit of the grand mean to the data or the lack-of-fit of the simple model. SSWithin, in contrast, measures the lack-of-fit of the group means to the data or the lack-of-fit of the full model.

In this example, SSTotal=25.28 and SSWithin=4.60. Because SSWithin is less than SSTotal that means that the full model that uses i fits the data better than the simple model that uses just μ.

However, we knew that this was going to happen as the full model always fits better. What we need now is a measure of how much better the full model fits or, equivalently, a measure of how much the lack-of-fit was reduced by using the full model rather than the simple model.

4.2.2 SSAmong

An useful property of SSTotal is that it "partitions" into two parts according to the following simple formula

$$\mathrm{SS}_{\mathrm{Total}} = \mathrm{SS}_{\mathrm{Within}} + \mathrm{SS}_{\mathrm{Among}}$$

This introduces a new quantity, SSAmong. A quick re-arrangement of the partitioning of SSTotal shows that

$$\mathrm{SS}_{\mathrm{Among}} = \mathrm{SS}_{\mathrm{Total}} - \mathrm{SS}_{\mathrm{Within}}$$

Thus, SSAmong records how much the lack-of-fit was reduced by using the full model rather than the simple model. In other words, SSAmong records how much "better" the full model fits the data than the simple model.

In our example, SSAmong=25.28-4.60=20.68. Thus, the residual SS from the simple model was reduced by 20.68 when the full model was used.

SSAmong is the **benefit** (i.e., reduction in lack-of-fit) of using the full rather than simple model

SSAmong can also be thought of in a different way. It can be algebraically shown that

$$\mathrm{SS}_{\mathrm{Among}} = \sum_{i=1}^{I} n_i \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2$$

Again, this looks complicated, but the main part to focus on is $\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$, which shows that SSAmong is primarily concerned with measuring the distance between the group means (i.e., $\bar{Y}_{i\cdot}$) and the grand mean (i.e., $\bar{Y}_{\cdot\cdot}$; Figure 4.4).

From the figure above, it is seen that SSAmong will increase as the group means become more different. In other words, SSAmong measures the **signal** in the data.

SSAmong is the **signal** (i.e., relative difference in group means) in the data

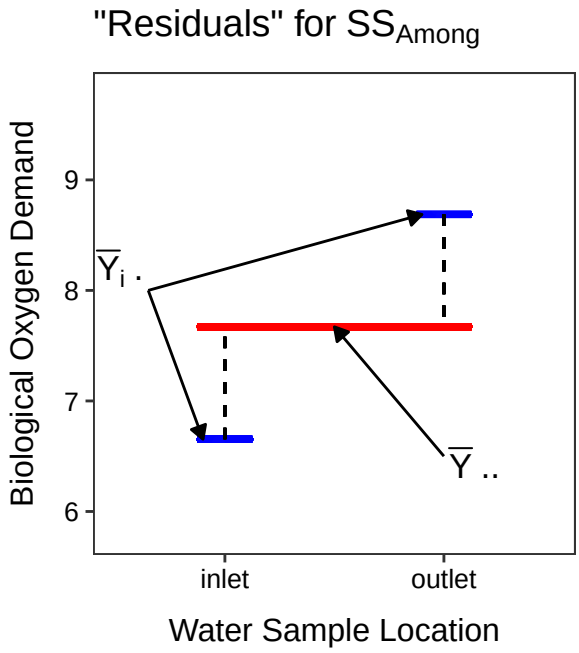
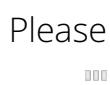


Figure 4.4: Mean biological oxygen demand versus sample location with the grand mean shown as a red horizontal segment and the group means shown as blue horizontal segments. Residuals between the group means and the grand mean are shown by black vertical dashed lines. The sum of these residuals scaled by the group sample sizes is SS_{Among}.

This can be seen in the interactive graphic below. You can adjust the amount of "signal" in the data by increasing or decreasing the difference between the group means and the grand mean. As you do this note how SSAmong (and SSTotal) change.



So, SSAmong is immensely useful – it is a measure of "benefit" that will be used in a "benefit-to-cost" ratio and it is the "signal" that will be used in a "signal-to-noise" ratio. These ratios are discussed further below. Next we discuss how to measure the "cost" of using the more complex full model.

4.3 Measuring Increase in Complexity

In this example, dfTotal=20-1 because there is one parameter (the grand mean) in the simple model, and dfWithin=20-2 because there are two parameters (the group means) in the full model. The full model uses more parameters and, thus, the residual degrees-of-freedom is reduced – there is a "cost" to using the full model over the simple model. We need a measure of this "cost".[3]

Interestingly dfTotal partitions in the same way as SSTotal; i.e.,

$$\mathrm{df}_{\mathrm{Total}} = \mathrm{df}_{\mathrm{Within}} + \mathrm{df}_{\mathrm{Among}}$$

[3]The "cost" is obviously 1 in this simple case.

This introduces another new quantity, dfAmong. A quick re-arrangement of the partitioning of dfTotal shows that

$$\mathrm{df}_{\mathrm{Among}} = \mathrm{df}_{\mathrm{Total}} - \mathrm{df}_{\mathrm{Within}}$$

In this case, dfAmong=19-18=1.

Thus, dfAmong is the degrees-of-freedom that were "lost" or "used" when the more complicated full model was used compared to the simpler simple model. The dfAmong is also the **difference in number of parameters** between the full and simple models. In other words, dfAmong is how much more complex (in terms of number of parameters) the full model is compared to the simple model. Thus, dfAmong measures the **cost** of using the full model rather than the simple model.

dfAmong is the extra **cost** (i.e., loss of df) from using the full rather than simple model

4.4 "Noise" Variances

MSTotal and MSWithin are measures of the variance [4] of **individuals** around the grand mean and group means, respectively. Thus, MSTotal measures the variance or "noise" around the full model, whereas MSWithin measures the variance or "noise" around the simple model.

MSTotal and MSWithin measure "noise" – i.e., variability of observations around a model

Before moving on to discuss MSAmong, it is worth noting that MSTotal is

$$\mathrm{MS}_{\mathrm{Total}} = \frac{\mathrm{SS}_{\mathrm{Total}}}{\mathrm{df}_{\mathrm{Total}}} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{..}\right)^2}{n-1}$$

Realizing that the double summation simply means to "sum across all individuals" it is seen that this is the variance (s^2) from your introductory statistics course. In other words it is just the variability of the individuals around a mean that ignores that there are groups.

MSTotal=s^2

[4] As discussed in Section 3.4, SS are not true variances until they are divided by their df and become mean-squares (MS).

Similarly, MSWithin is

$$\text{MS}_{\text{Within}} = \frac{\text{SS}_{\text{Within}}}{\text{df}_{\text{Within}}} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2}{\sum_{i=1}^{I} n_i - I}$$

It is not hard to show algebraically (and for just two groups) that the numerator is $n_1 s_1^2 + n_2 s_2^2$ and that the denominator is $n_1 + n_2 - 2$. This numerator and denominator are then simply the pooled sample variance (s_p^2) from the 2-sample t-test. Thus, MSWithin with two groups is the same as s_p^2 from the 2-sample t-test.

MSWithin=s_p^2

4.5 "Signal" Variance (Benefit-to-Cost)

Of course SSAmong divided by dfAmong will be MSAmong. However, while MSAmong is still a variance, it has a very different interpretation.

MSAmong is NOT a variance of *individuals*, rather it is a variance of *sample means*. Sample means can vary (i.e., not be equal) for two reasons – purely due to random sampling variability (i.e., the population means are not different) or the population means really do differ such that the sample means differ. In other words, MSAmong – the variance among means – is a combination of "noise" and "signal." Our goal (next) is to disentangle these two reasons for why the sample means differ to determine if there is a real "signal" or not.

Additionally, MSAmong is a ratio of the "benefit" (i.e., SSAmong) to the "cost" (i.e., dfAmong) of using the full model over the simple model. So MSAmong scales the benefit to the cost of using the full model.

4.6 Ratio of Variances (Signal-to-Noise)

From the above discussion we have a measure of potential "signal" in MSAmong and actual "noise" around the full model (the model representing the "signal") in MSWithin. The ratio of this "signal" to "noise" is called an F test statistic; i.e.,

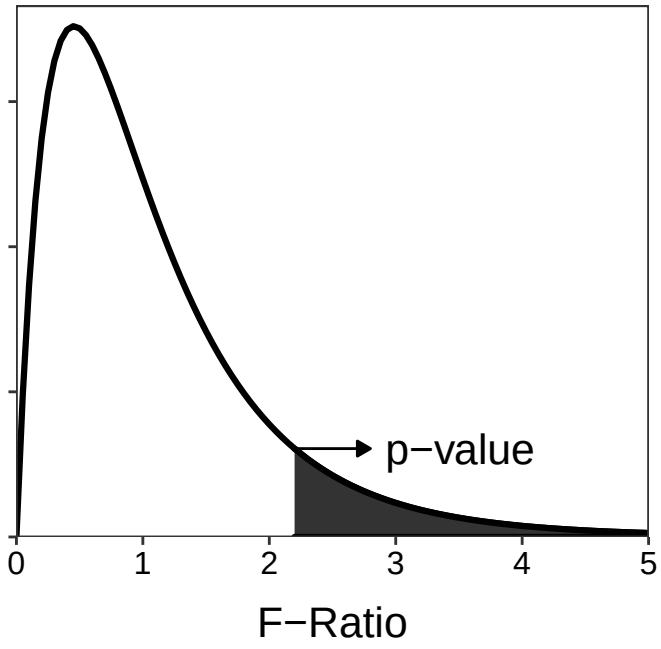
$$\text{F} = \frac{\text{MS}_{\text{Among}}}{\text{MS}_{\text{Within}}} = \frac{\text{Signal}}{\text{Noise}} = \frac{\text{Variance Explained by Full Model}}{\text{Variance Unexplained by Full Model}}$$

If the F-ratio is "large," then a great deal more variability was explained (i.e., more "signal") than was unexplained by the full model (i.e., "less noise") and one would conclude that the full model fits the data significantly better than the simple model, even considering the increased complexity of the full model.

The question now becomes "when is the F-ratio considered large enough to reject the simple model and conclude that the full model is significantly better?" This question can by comparing the F-ratio test statistic to an F-distribution.

An F-distribution [5] is right-skewed, with the exact shape of the distribution dictated by two separate degrees-of-freedom – called the numerator and denominator degrees-of-freedom, respectively. The numerator df is equal to the df used in MSAmong, whereas the denominator df is equal to the df used in MSWithin. The p-value is always computed as the area under the F-distribution curve to the right of the observed F-ratio test statistic.[6]

The p-value is always computed to the **right** on an F-distribution.



From this it can be seen that a small p-value comes from a large F-ratio, which comes from a large MSAmong relative to MSWithin, which means both that

[5] An F-distribution occurs whenever the ratio of two variances is calculated.

[6] If the F-ratio is computed by hand, then `distrib()` with `distrib="f"`, `df1=`, `df2=`, and `lower.tail=FALSE` may be used to calculate the corresponding p-value.

the full model explains more variability than is left unexplained and the "signal" is much greater than the "noise", which means that the full model does fit significantly better than the simple model (even given the increased complexity), and, thus, the means are indeed different, which is what we would conclude from a small p-value. This cascade of measures can be explored with the dynamic graphic below.

Explore F-Ratio and p-value

The graphic below shows obserations relative to a grand mean (horizontal red bar) and two group means (horizontal blue lines). Residuals from group means are shown with vertical blue dashed lines and the difference between the group means and the grand mean are shown with vertical red dashed lines. You can alter the difference between the group means and the grand mean with the slider bar to see how this affects SSTotal and its two components -- SSWithin (height of blue bar on right) and SSAmong (height of red bar on right) -- and the F-ratio test statistic and corresponding p-value.

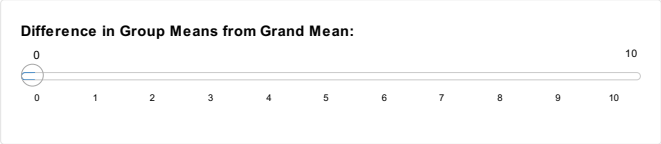


Table 4.2: An ANOVA table for biological oxygen demand measurements at two locations of the aquaculture facility. Note that the "Total" row is not shown.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
src	1	20.6756	20.6756	80.8912	0
Residuals	18	4.6008	0.2556		

4.7 ANOVA Table

The degrees-of-freedom (df), sum-of-squares (SS), mean-squares (MS), F-ratio test statistic (F), and corresponding p-value are summarized in what is called an **analysis of variance (ANOVA) table**.[7] The ANOVA table contains rows that correspond to the different measures discussed above: among,[8] within,[9] and total. The df and SS are shown for each source, but the MS is shown only for the within and among sources because MSAmong+MSWithin MSTotal.

SS and df partition, but MS do not! Do not add MSAmong and MSWithin to get MSTotal, instead divide SSTotal by dfTotal.

An ANOVA table for the BOD measurements at the inlet and outlet sources to the aquaculture facility is in Table 4.2. Note that R does not show the total row that most softwares do.

These results indicate that H0 should be rejected (i.e., F-test p-value <0.00005). Thus, the full model fits the data significantly better than the simple model even given the difference in complexity between the two models and sampling variability. Therefore, there is a significant difference in mean BOD between the two locations.

In addition to the primary objective of comparing the full and simple models, several items of interest can be identified from an ANOVA table. Using the table above as an example, note the following items:

- The variance within groups is equal to MSWithin (e.g., MSResiduals=0.2556 in this case). This is s_p^2 from the two-sample t-test (because there are only two groups here).

[7] An ANOVA table does not necessarily mean that an "analysis of variance" method was used. It turns out that all general linear models are summarized with an ANOVA table, regardless of whether a one- or two-way ANOVA method was used.

[8] Labeled as the factor variable in most statistical software packages including R – that variable was called `src` in this example.

[9] Labeled as residuals in R and error in other statistical software packages.

- The common variance about the mean (s^2) is given by MSTotal (e.g., $= \frac{20.6756+4.6008}{1+18}$=1.3303).

4.8 Two-Sample t-Test Revisited: Using Linear Models

The models for a two-sample t-test can be fit and assessed with `lm()`. This function requires the same type of formula for its first argument – `response~factor` – and a data.frame in the `data=` argument as described for `t.test()` in Section 2.2. The results of `lm()` should be assigned to an object so that specific results can be selectively extracted from it. For example, the ANOVA table results are extracted from the `lm()` object with `anova()`. In addition, coefficient results[10] can be extracted with `coef()`, `confint()`, and `summary()`. Note that I like to "column-bind" the coefficients and confidence intervals together for a more succinct representation.

```
aqua.lm <- lm(BOD~src,data=aqua)
anova(aqua.lm)
```

```
#R>   Analysis of Variance Table
#R>
#R>   Response: BOD
#R>             Df  Sum Sq Mean Sq F value    Pr(>F)
#R>   src        1 20.6756 20.6756  80.891 4.449e-08
#R>   Residuals 18  4.6008  0.2556
```

```
cbind(ests=coef(aqua.lm),confint(aqua.lm))
```

```
#R>                   ests    2.5 %   97.5 %
#R>   (Intercept) 6.6538 6.317917 6.989683
#R>   srcoutlet   2.0335 1.558489 2.508511
```

From these results, note:

- The p-value in the ANOVA table is the same as that computed from `t.test()`.
- The coefficient for `srcoutlet` is the same as the difference in the group means computed with `t.test()`.
- The F test statistics in the ANOVA table equals the square of the t test statistic from `t.test()`. This is because an F with 1 numerator and v denominator df exactly equals the square of a t with v df.

[10]The coefficient results will be discussed in more detail in Module 5.

Thus, the exact same results for a two-sample t-test are obtained whether the analysis is completed in the "traditional" manner (i.e., with `t.test()`) or with competing models (i.e., using `lm()`). This concept will be extended in subsequent modules.

4.9 One More Look at MS and F-test

Recall from your introductory statistics course that a sampling distribution is the distribution of a statistic from all possible samples. For example, the Central Limit Theorem states that the distribution of sample means is approximately normal, centered on , with a standard error of $\frac{\sigma}{\sqrt{n}}$ as long as assumptions about the sample size are met. Further recall that the sampling distribution of the sample means is centered on because the sample mean is an unbiased estimator of . Similarly, it is also known that the center of the sampling distribution of s^2 is equal to σ^2 because s^2 is an unbiased estimate of σ^2.

MSWithin and MSAmong are statistics just as $\bar{x}$ and s^2 are statistics. Thus, MSWithin and MSAmong are subject to sampling variability and have sampling distributions. It can be shown[11] that the center of the sampling distribution of MSWithin is σ^2 and the center of the sampling distribution of MSAmong is

$$\sigma^2 + \frac{1}{I-1}\sum_{i=1}^{I} n_i \left(\mu_i - \mu\right)^2$$

Thus, MSAmong consists of two "sources" of variability. The first source (σ^2) is the natural variability that exists among individuals. The second source $\left(\frac{1}{I-1}\sum_{i=1}^{I} n_i \left(\mu_i - \mu\right)^2\right)$ is related to differences among the group means. Therefore, if the group means are all equal – i.e., μ_1=μ_2= $\cdots$ = μ_I = μ – then the second source of variability is equal to zero and MSAmong will equal MSWithin. As soon as the groups begin to differ, the second source of variability will be greater than 0 and MSAmong will be greater than MSWithin.

From this, it follows that if the null hypothesis of equal population means is true (i.e., one mean fits all groups), then the center of the sampling distribution of both MSWithin and MSAmong is σ^2. Therefore, if the null hypothesis is true, then the F test-statistic is expected to be equal to 1, on average, which will always result in a large p-value and a DNR H0 conclusion. However, if the null hypothesis is false (i.e., separate means are needed for all groups), then the center of the sampling distribution of MSWithin is σ^2 but the center of the sampling distribution of MSAmong is σ^2 + "something", where the "something" is greater than 0 and gets larger as the means become "more different." Thus, if the null hypothesis is false then the F test-statistic is expected to be greater than

[11] This derivation is beyond the scope of this course.

1 and will get larger as the null hypothesis gets "more false." This analysis of sampling distribution theory illustrates once again that (1) MSAmong consists of multiple sources of variability and (2) "large" values of the F test-statistic indicate that the null hypothesis is incorrect.

ONE-WAY ANOVA

Chapter 5

One-Way Foundations

Many studies, including the following examples, result in the comparison of means from more than two independent populations.

- Determine if the mean volume of white blood cells of Virginia opossums (*Didelphis virginiana*) differed by season in the same year (Woods and Hellgren 2003).
- Determine if the mean frequency of occurrence of badgers (*Meles meles*) in plots differs between plots at different locations (Virgos and Casanovas 1999).
- Test for differences in the mean total richness of macroinvertebrates between the three zones of a river (Grubbs and Taylor 2004).
- Test if the mean mass of porcupines (*Erithizon dorsatum*) differs among months of summer (Sweitzer and Berger 1992).
- Test if the mean clutch size of spiders differs among three types of parental care categories (Simpson 1995).
- Determine if the mean age of harvested deer (*Odocoelius virginianus*) differs among 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 2-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**) is simply an extension of a 2-sample t-test and can be used for each of these situations.[1]

[1]This and the next several modules depends heavily on the foundational material in Modules 1-4, especially the concepts of simple and full models; "signal" and "noise"; variances explained and unexplained; and SS, MS, F, and p-values.

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

In this module, we examine the immunoglobulin[2] measurements of opossums (`imm`) during three months of the same year (`season`). The data are loaded into R and a subset of rows is shown below.

```
opp <- read.csv("Opossums.csv")
```

```
head(opp)
```

```
#R>       imm season
#R>   1 0.640    feb
#R>   2 0.680    feb
#R>   3 0.731    feb
#R>   4 0.587    feb
#R>   5 0.668    feb
#R>   6 0.613    feb
```

Data must be stacked!!

5.1 Analytical Foundation

The generic null hypothesis for a one-way ANOVA is

$$\mathrm{H}_0 : \mu_1 = \mu_2 = ... = \mu_I$$

where I is the total number of groups or populations.[3] 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

$$\mathrm{H}_\mathrm{A} : \text{At least one pair of means is different}$$

A rejection of H0 in favor of HA 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 Module 6.

[2] Any of a class of proteins present in the serum and cells of the immune system, which function as antibodies.

[3] From this, it is evident that the one-way ANOVA is a direct extension of the 2-sample t-test.

Rejecting H0 **just** means that **some** group means differ.

The simple ($Y_{ij} = \mu + \epsilon_{ij}$) and full ($Y_{ij} = \mu_i + \epsilon_{ij}$) models for the one-way ANOVA are the same as those for the 2-sample t-test, except that there are I>2 means in the full model. Thus, SStotal, SSwithin, and SSamong are computed using the same formulae shown in Module 4, except to again note that I>2. The degrees-of-freedom are also computed similarly – i.e., dfWithin=n-I and dfAmong=I-1. The MS, F, and p-value are also computed in the same way.[4]

A 2-Sample t-Test is simply a special case of a One-Way ANOVA.

Figure 5.1 is a visual for simple and full models and residuals from each. Note the similarity with figures from the Module 4, except that there are three group means here.

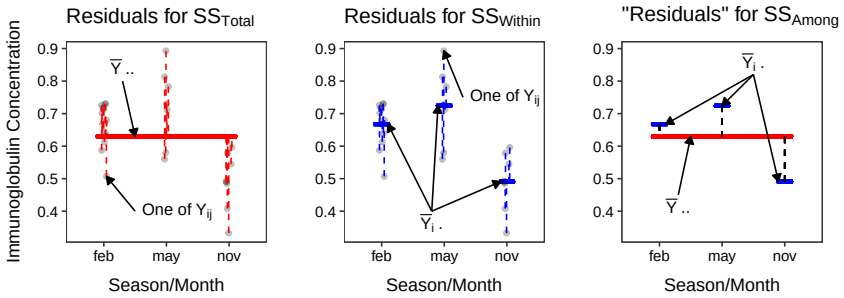


Figure 5.1: Immunoglobulin concentrations versus season (or month) of capture for New Zealand opposums. The grand mean is shown by a red horizontal segment, group means are shown by blue horizontal segments, residuals from the grand mean are red vertical dashed lines, residuals from the groups means are blue vertical dashed lines, and differnces between the group means and the grand mean are black vertical dashed lines.

An ANOVA table (Table 5.1) is use to display the results from a one-way ANOVA, because the one-way ANOVA is simply a comparison of two models.

In addition to the usual meanings attached to MSAmong, MSAmong, and MSTotal,[5] the following can be discerned from this ANOVA table.

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

[5]Note that MSTotal **must** be computed from SSTotal and dfTotal and not by summing MSAmong an MSWithin.

Table 5.1: An ANOVA table for immunoglobuling concentration by season (or month) for New Zealand opossums. Note that the "Total" row is not shown.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
season	2	0.2340	0.1170	14.4486	1e-04
Residuals	24	0.1944	0.0081		

- dfAmong=2 and because dfAmong=I-1, then I=3. This confirms that there are three groups in this analysis.
- dfTotal=dfAmong+dfWithin=2+24=26. Because dfTotal=n-1, then n=27. This shows that there are 27 individuals in this analysis.
- There is a significant difference in the mean immunoglobulin values among the three months because the p-value=0.0001< .

5.2 One-Way ANOVA in R

The models for a one-way ANOVA are fit and assessed with `lm()` exactly as described for a 2-sample t-test in Section 4.8. As a reminder, a formula of `response~factor` is the first argument and a data.frame is given in `data=` in `lm()`, the results of `lm()` should be assigned to an object, and the ANOVA table is extracted with `anova()`.

The `lm()` code is the same for a 2-Sample t-Test and a One-Way ANOVA.

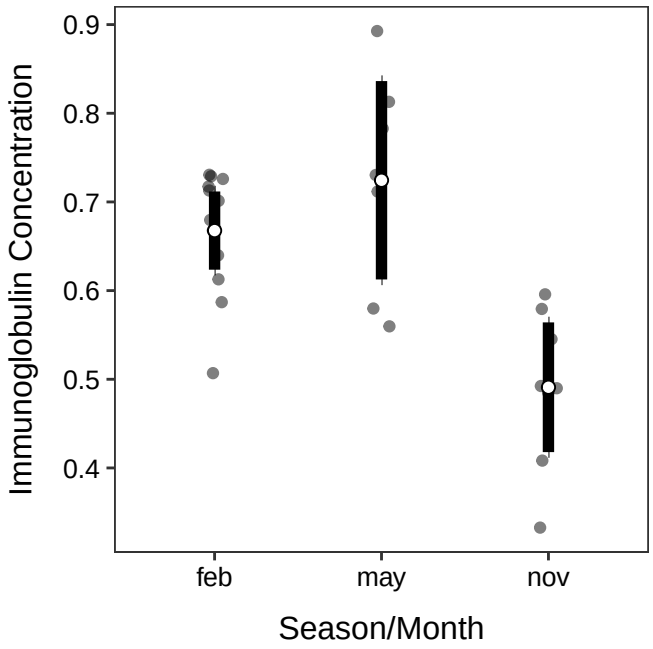
```
lm1 <- lm(imm~season,data=opp)
anova(lm1)
```

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

A graphic that illustrates the mean immunoglobulin value with 95% confidence intervals for each month is constructed below (as shown in Section 2.2).

```
ggplot(data=opp,mapping=aes(x=season,y=imm)) +
  geom_jitter(alpha=0.5,width=0.05) +
  stat_summary(fun.data=mean_cl_normal,geom="errorbar",size=2,width=0) +
  stat_summary(fun=mean,geom="point",pch=21,fill="white",size=2) +
```

```
labs(y="Immunoglobulin Concentration",x="Season/Month") +
theme_NCStats()
```



Chapter 6

One-Way Multiple Comparisons

A significant result (i.e., reject H0) in a one-way ANOVA indicates that the means of at least one pair of groups differ. It is not yet known whether all means differ, all but two means differ, only one pair of means differ, or any other possible combination of differences. Thus, specific follow-up analyses to a significant one-way ANOVA are needed to identify which pairs of means are significantly different.

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

6.1 Multiple Comparison Problem

The most obvious solution to identify which pairs of means differ is to perform a 2-sample t-test for each pair of groups. Unfortunately, this seemingly simple answer has at least two major problems. First, the number of 2-sample t-tests needed increases dramatically with increasing numbers of groups. 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. Of these two issues, the second is much more problematic and needs to be better understood.

In any one comparison of two means the probability of incorrectly concluding that the means are different when they are actually not different is . This incorrect conclusion is called a **pairwise Type I error** because it relates to

only one comparison of a pair of means.

In a situation with three (I=3) groups (say A, B, C) then there are three pairwise comparisons (k=3) to be made (A to B, A to C, and B to C). A pairwise error could be made on any of these three tests. Making a Type I error on *at least one* of these multiple pairwise tests is called an **experiment-wise Type I error** because it involves all pairwise comparisons in the experiment at hand.

It is important that you notice *at least* in the definition of the experiment-wise error rate. For example, 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!!

A Type I error is rejecting H0 when H0 is actually true. In a two-sample t-test, a Type I error is concluding that two means are significantly different when they are not different.

Pairwise error rate: The probability of a Type I error in a single comparison of two means. Sometimes called an comparison-, individual-, or test-wise error.

Experiment-wise error rate: The probability of *at least one* Type I error in a set of comparisons of two means. Sometimes called the family-wise error.

Figure 6.1 demonstrates the two issues related to multiple comparisons. First, the x-axis labels show how the number of pairwise comparisons (k) increases quickly with increasing number of groups (I) in the study. For example, six groups (I=6) is not a complicated study, but it results in fifteen pairwise comparisons (k=15). More importantly the line and point labels in the figure show how the experiment-wise error rate increases quickly and dramatically with increasing number of groups. For example, the experiment-wise error rate for six (I=6) groups is over 0.50.[1] Thus, it is nearly a coin flip that at least one error will be made in all paired comparisons among six groups. Making an error more than 50% of the time in such a simple study is not acceptable and must be corrected.

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

[1] Using α=0.05

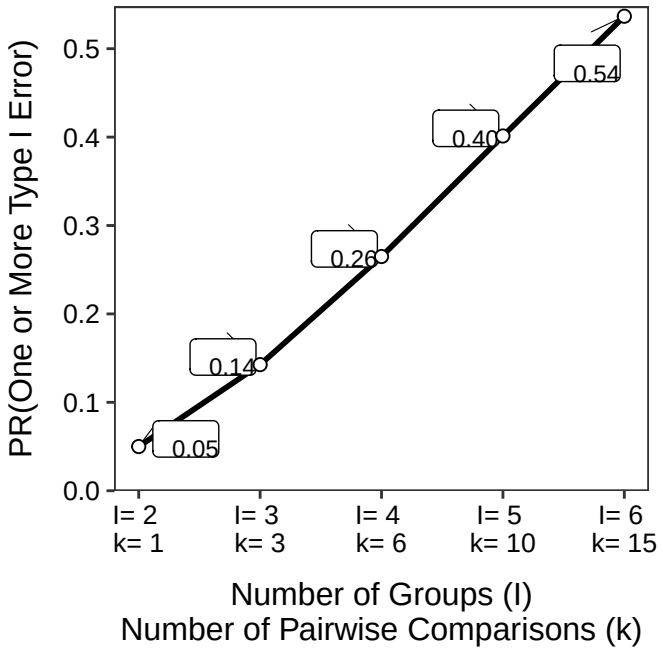


Figure 6.1: Relationship between the number of groups (I) in an analysis, the number of pairs of means that would need to be tested (k), and the probability of making one or more Type I errors in all comparisons. Note that alpha=0.05.

6.2 Correction Methods

There are many procedures designed to attempt to control experiment-wise error rate at a desired level (usually). You will here a variety of names like Tukey's HSD, Bonferroni's adjustment, Sidak's method, and Scheffe's method.[2] For simplicity, only the Tukey-Kramer honestly significantly different (i.e., Tukey's HSD or Tukey's) method will be used here.

As simplistically as possible, Tukey's test computes the t test statistic for each pair of means as if conducting a 2-sample t-test. However, this test statistic is compared to a "Studentized range" rather than a t distribution to compute the p-value. These "adjusted" p-values are then simply compared to to make a decision about the means of each pair. The net result of this modification however is that the experiment-wise error rate across all comparisions is controlled at the desired level when the group sample sizes are equal and is slightly conservative when the group sample sizes are different.

6.3 Multiple Comparisons in R

Tukey's procedure should only be implemented if multiple comparisons are needed!! In other words, only use this method following a significant One-Way ANOVA result; i.e., H0 was rejected such that it appears that there is some difference among group means. Therefore, a One-Way ANOVA must be performed first as described in Section 5.2.

The ANOVA table from the analysis of immunoglobulin levels in opossums across seasons that was begun in the Module 5 is shown below.

```
lm1 <- lm(imm~season,data=opp)
anova(lm1)
```

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

Once again, there appears to be some significant difference in the mean immunoglobulin values among the three months (0.0001<). Thus, a multiple comparisons procedure is warranted here to identify exaclty which pairs of means differ.

There are a number of functions and packages in R for computing Tukey's

[2] See here for a short list of methods.

multiple comparisons. I prefer to use functions in the `emmeans` package because those functions will generalize to other methods, some of which we will use in other modules and some of which you may use in more advanced statistics courses. The `emmeans` package must be attached with `library()` before its functions can be used.

The `emmeans` package must be attached with `library` to perform Tukey's procedure.

```
library(emmeans)
```

Tukey's procedure is computed with a two-step process. First, use `emmeans()` with the `lm()` object as the first argument and a `specs=` argument with `pairwise~` followed by the name of the variable that identifies the groups. The results from this function should be saved to an object.

```
mc <- emmeans(lm1,specs=pairwise~season)
```

That saved object is then the first argument to `summary()`, which also uses `infer=TRUE`. This again should be saved to an object.

```
( mcsum <- summary(mc,infer=TRUE) )
```

```
#R>   $emmeans
#R>    season emmean     SE df lower.CL upper.CL t.ratio p.value
#R>    feb     0.668 0.0260 24    0.614    0.721 25.702  <.0001
#R>    may     0.724 0.0340 24    0.654    0.795 21.299  <.0001
#R>    nov     0.491 0.0318 24    0.425    0.557 15.433  <.0001
#R>
#R>   Confidence level used: 0.95
#R>
#R>   $contrasts
#R>    contrast  estimate     SE df lower.CL upper.CL t.ratio p.value
#R>    feb - may  -0.0568 0.0428 24  -0.1636   0.0501 -1.326  0.3948
#R>    feb - nov   0.1767 0.0411 24   0.0741   0.2792  4.301  0.0007
#R>    may - nov   0.2334 0.0466 24   0.1171   0.3497  5.012  0.0001
#R>
#R>   Confidence level used: 0.95
#R>   Conf-level adjustment: tukey method for comparing a family of 3 estimates
#R>   P value adjustment: tukey method for comparing a family of 3 estimates
```

The results are in two "sections" labeled as `$emmeans` and `$contrasts`.

The `$contrasts` section contains the actual Tukey's test for each pair of means. In these results the difference in group sample means is under `estimate`, a 95% confidence interval for the **difference** in means is under `lower.CL` and

`upper.CL`, and a p-value for testing that the difference in group population means is 0 is under `p.value`. For example, the difference in group **sample** mean immunoglobulin between February and May is -0.0568, but the p-value suggests that the **population** mean immunoglobulin does not differ between February and May (p=0.3948). In contrast, it appears that the population mean immunoglobulin for opossums in November differed from both those in Feb (p=0.0007) and those in May (p=0.0001).

The **difference** of group means with 95% confidence intervals and p-values are shown in the `$contrasts` section of the results.

The `$emmeans` section contains the group sample means under `emmean` with 95% confidence intervals under `lower.CL` and `upper.CL`. For example, the sample mean immunoglobulin level for opossums in February was 0.668, with a 95% confidence interval from 0.614 to 0.721. The `t.ratio` and `p.value` in this section tests if the group population mean is different than 0. These tests are not often of interest and can largely be ignored.

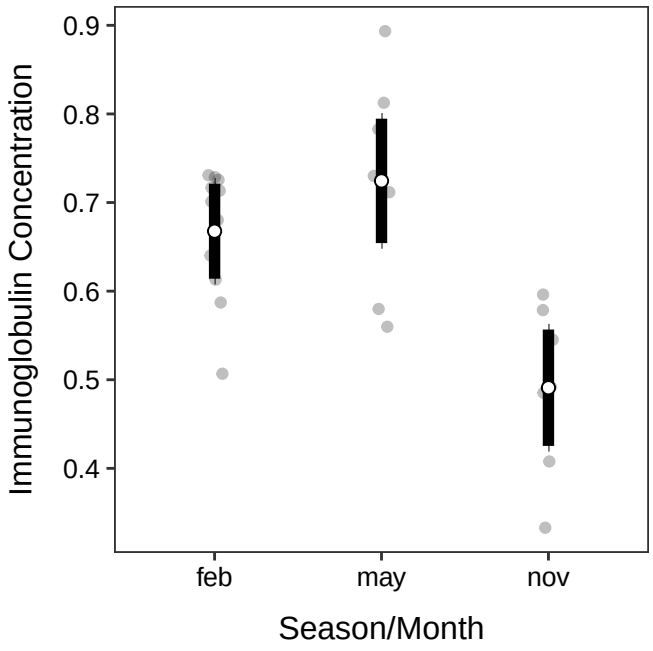
The group means with 95% confidence intervals are shown in the `$emmeans` section of the results.

A plot of group means with 95% confidence intervals using the results in `$emmeans` is slightly different than shown in Sections 4.8 and 5.2 because the raw data and the means with their confidence intervals are in separate data frames. While this method is slightly more complicated, it will generalize to a wider variety of situationst throughout the course.

The `data=` and `mapping=aes()` arguments are not included in the initial `ggplot()` because we will be drawing variables from two data frames. Thus, `geom_jitter()` below adds the raw data to the plot, jittered to decrease overlap; `geom_errorbar()` creates the error bars from the `$emmeans` object, and `geom_point()` simply overlays the mean from the `$emmeans` object. Note that in the code below you would only need to modify the two `data=` arguments, the three `x=` arguments (to the grouping variables), and the one `y=` argument in `geom_jitter()` (to the response variable).

```
ggplot() +
  geom_jitter(data=opp,mapping=aes(x=season,y=imm),
              alpha=0.25,width=0.05) +
  geom_errorbar(data=mcsum$emmeans,
                mapping=aes(x=season,ymin=lower.CL,ymax=upper.CL),
                size=2,width=0) +
  geom_point(data=mcsum$emmeans,mapping=aes(x=season,y=emmean),
             size=2,pch=21,fill="white") +
  labs(y="Immunoglobulin Concentration",x="Season/Month") +
```

```
theme_NCStats()
```



Chapter 7

One-Way Assumptions

As with most statistical methods, the One-Way ANOVA requires that four assumptions be met so that the calculations made in Modules 5 and 6 mean what we said they would mean. The four assumptions for a One-Way ANOVA are:[1]

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

Each of these assumptions in more detail below.

7.1 Independence

In a One-Way ANOVA the individuals **must be** independent both within and among groups. In other words, there must be no connection between individuals within a group or between individuals in one group and individuals in other groups.

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

A lack of independence may include applying multiple treatments to the same individual, having related individuals either within the same group or specifically spread across the groups, or having individuals that are not separated in space or time. Below are examples where there was a *lack of independence*.

[1] Note that the first three are the same three assumptions you learned for a 2-Sample t-test.

- Researchers measured the self-esteem at three time points – beginning, middle, and end – for 10 people following a specific diet. They wanted to determine if self-esteem increased over the time that individuals were on the diet. This example illustrates a lack of *among-group* independence because the same 10 people were in each of the "groups" (i.e., beginning, middle, and end time period).
- Zoo keepers were interested in whether the activity rate of lions differed by time of day. For this study, they recorded the activity rate of all five lions at the same random times in the morning, afternoon, evening, and night across several days. This example illustrates a lack of *among-group* independence because the same lions were recorded in each period. This also illustrates a lack of *within-group* independence because the activity rates were recorded at the same times for each lion. Thus, the lions may be affecting each others' activity rates. For example, if one lion gets up to roam around then the other lions may be more likely to get up to roam around as well.
- Researchers with the LoonWatch program wanted to determine if mean density of loons differs among Bayfield, Ashland, and Iron counties. For this they asked local volunteers to record the number of loons they observed on several lakes during the same weekend in June. This example illustrates a lack of *within-group* independence as different observers were used in each county. It is possible that the observers in one county are more adept at observing loons on their lakes (for whatever reason – they know their lakes better, their lakes are smaller, they are more motivated, they spend more time).

There are methods to detect violations of the independence assumption if the assumption is related to time (e.g., the first situation above), but for most other situations a violation is only 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.[2] In other words, designing data collections with independence among individuals is critical and needs to be ascertained before the data are collected.

Independence is generally assessed by considering how the individuals were obtained.

In this course, the data will have already been collected for you and, at times, the description of that data collection may be sparse. To address independence you will be asked to explain why you think dependencies do not exist in the data collection. This may take several sentences. Examples will be provided below and in future analyses.

[2]Such methods may include repeated measures ANOVA, mixed-models, and hierarchical models.

7.2 Equal Variances

The variances among groups must be equal because the estimate of MSWithin is based on pooling estimates across the groups. In other words, if the variances among each group are equal, then the variance (or MS) for each group is an estimate of the overall MSWithin. If the variances are equal across groups then combining the variances from each group provides a robust estimate of the overall variance within groups.

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

The assumption of equal variances can be tested with Levene's homogeneity of variances test.[3] The hypotheses tested by Levene's test are

$$\begin{aligned} &\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2 \\ &\mathrm{H}_\mathrm{A} : \text{At least one pair of variances differ} \end{aligned}$$

Thus, a p-value less than means that the variances are not equal and the assumption of the one-way ANOVA has not been met.[4]

The equality of variances may be visually examined with a boxplot of full model residuals[5] by group. If the "boxes" on this boxplot are not roughly the same, then the equal variances assumption may be violated. I will usually examine the boxplots rather than use a Levene's Test when the sample size is very large because Levene's test can be hyper-sensitive with large samples sizes (i.e., reject H0 of equal variances when the variances are not practically different).

7.3 Normality

The normality of residuals WITHIN each group is difficult to test because there may be many groups being considered or relatively few individuals in each group. Thus the normality of all residuals taken as a whole is often tested. As most linear models are resilient to slight departures from normality, it is thought that

[3] 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.

[4] Methods for "working around" this assumption violation are discussed in Module 8.

[5] Recall that these are the vertical differences between observations and their group mean.

if all of the residuals as a whole appear approximately normal then the residuals within each group are likely "normal enough."

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

Normality is often tested by simply viewing a histogram of residuals or a so-called Q-Q plot. For an adequate sample size, a histogram that is not strongly skewed is probably adequate for a One-Way ANOVA.

The normality of residuals may also be tested with the Anderson-Darling Normality Test.[6] The hypotheses for this test are

$$H_0 : \text{Residuals are normally distributed}$$
$$H_A : \text{Residuals are not normally distributed}$$

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

The results of an Anderson-Darling test should be interpreted cautiously for both small and very large sample sizes. At small sample sizes, the distribution would need to be wildly non-normal for the Anderson-Darling Test to suggest that it is not normal. At very large sample sizes, very small and insubstantial differences from normality may result in the test indicating that the distribution is not normal. Thus, it is important to always examine the histogram of residuals to decide whether this assumption is adequately met or not.

7.4 No Outliers

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) or 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.

[6] 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 course.

A one-way ANOVA is very sensitive to outliers.

Outliers may be detected by visual examination of a histogram of residuals.

Potential outliers can be more objectively detected with externally Studentized residuals,[7] which essentially measure how many standard deviations an individual is from its group mean. Studentized residuals follow a t-distribution with dfWithin-1 degrees-of-freedom.[8] A p-value for testing whether an individual residual is an outlier or not is calculated by converting the Studentized residual to a two-tailed p-value using a t-distribution. As these p-values are computed for each residual, this process suffers from the "multiple comparison problem" (see Section **??**). Thus, the p-values use a Bonferroni method[9] to correct for multiple comparisons so that the likelihood of mistakingly identifying an outlier is controlled at a desirable level. If the Bonferroni adjusted p-value for the most extreme residual is less than , then that individual is considered to be a significant outlier and should be flagged for further inspection as described above.

7.5 Testing Assumptions in R

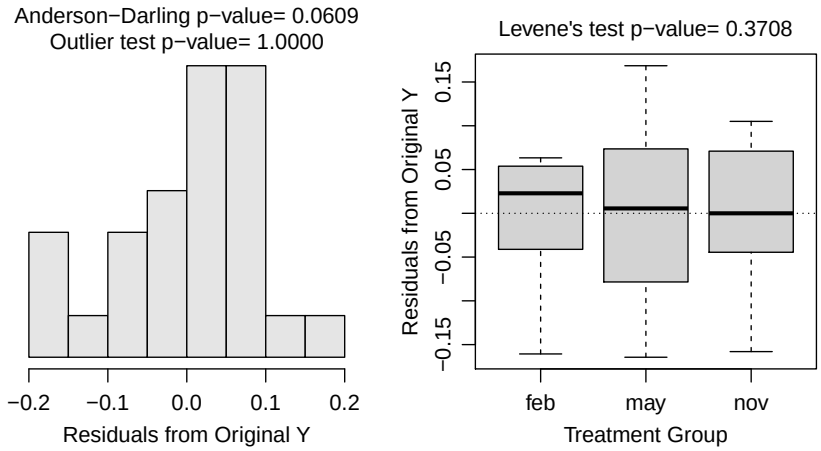
All plots and tests of assumptions can be completed by submitting the saved `lm` object from when the One-Way ANOVA was computed to `assumptionCheck()`. For example, the code below fits the One-Way ANOVA for testing if the mean immunoglobulin levels of New Zealand opossums differs among seasons (the `opp` data frame was created in Module 6) and then performs the calculations needed to check the assumptions.

```
lm1 <- lm(imm~season,data=opp)
assumptionCheck(lm1)
```

[7] A residual divided by the standard deviation of the residual, where the standard deviation is computed with that individual removed.

[8] The extra one is subtracted because the individual residual is not included in the calculation of the standard deviation of residuals.

[9] An adjusted p-value is computed by multiplying the original p-value by the number of comparisons made (in this case n).



For a One-Way ANOVA, `assumptionCheck()` produces a histogram of residuals with the Anderson-Darling and outlier test p-values on the left and a boxplot of residuals for each group with the Levene's Test p-value on the right.

In this case the boxes on the boxplot are similarly sized and the Levene's test p-value (=0.3708) is greater than suggesting that the group variances are equal. The histogram of residuals is difficult to assess because the sample size is so small, but it does not appear strongly skewed and the Anderson-Darling p-value (=0.0609) is (barely) greater than , which weakly suggests that the residuals are normally distributed. The histogram does not show any "odd" individuals and the outlier test p-value (=0.0402) is greater than which suggests that there are not any significant outliers in these data. Thus, the three assumptions that can be tested with the data all appear to be met.

The independence assumption cannot be assessed from the data and must be reasoned through. While there is not much information about this study, I will assume between group independence as there is no suggestion that the same opossums were sampled in each of the three seasons (i.e., no indication that they were tagged or otherwise individually identified). This is particularly clear because the sample size differs across seasons (see table below). I will also assume that there is within-group independence because there is no evidence that the opossums within any given season were somehow related or connected.

```
xtabs(~season,data=opp)
```

```
#R>   season
#R>   feb may nov
#R>    12   7   8
```

Chapter 8

One-Way Transformations

As discussed in Module 7 a One-Way ANOVA depends on four assumptions being met. If those assumptions are not met, then the results of the one-way ANOVA are invalid. Fortunately, violations of the equality of variances and normality assumptions can often be addressed by transforming the quantitative response variable to a scale where the assumptions are met. For example it is common use the natural log of the response variable rather than the response variable on its original scale.

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.

Besides the obvious reason related to assumption violations, Fox (1997) gave four arguments for why data that is skewed or has unequal variances 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.

Identification of an appropriate transformation and understanding the resultant output is the focus of this module.

8.1 Power Transformations

With power transformations, the response variable is transformed by raising it to a particular power, λ, i.e., Y^{λ} (Table 8.1).

Table 8.1: Common power transformations in ANOVAs.

Power	Name	Formula	R Code
λ=1	–	$Y^1 = Y$	–
λ=0.5	Square Root	$Y^{0.5} = \sqrt{Y}$	`df$newvar <- sqrt(df$oldvar)`
λ=0.33	Cube Root	$Y^{0.33} = \sqrt[3]{Y}$	`df$newvar <- df$oldvar^(1/3)`
λ=0.25	Fourth Root	$Y^{0.25} = \sqrt[4]{Y}$	`df$newvar <- df$oldvar^(1/4)`
λ=0	Natural log	$log(Y)$	`df$newvar <- log(df$oldvar)`
λ=-1	Inverse	$Y^{-1} = \frac{1}{Y}$	`df$newvar <- 1/df$oldvar`

Each transformation in Table 8.1 "spreads out" small values and "draws in" large values in a distribution. For example, in Figure 8.1 the log function is shown as the black curved line, the original histogram of strongly right-skewed data is shown upside-down below the x-axis, and the histogram following the log-transformation is shown sideways left of the y-axis. Two "small" values are shown in red on the x-axis. The log-transformation of these two points is shown by following the two vertical lines from these points to the black log function and then moving horizontally to the y-axis. From this it is seen that these two values that were relatively close together are more spread out after the transformation. Conversely two "large" values are shown in blue on the x-axis. Following the same process it is seen that these two points are closer together following the log transformation.

As seen in Figure 8.1, applying the log transformation to all values in the original strongly right-skewed distribution resulted in a distribution that was approximately normal. All transformations may not achieve normality. For example, the same process shown in Figure 8.1 is repeated in Figure 8.2 for a square-root transformation. A comparison of the two figures shows that the square-root transformation function is less curved, it spreads out relatively small values less and draws in relative larger values less, and it does not transform the original strongly right-skewed distribution to an approximately normal distribution.

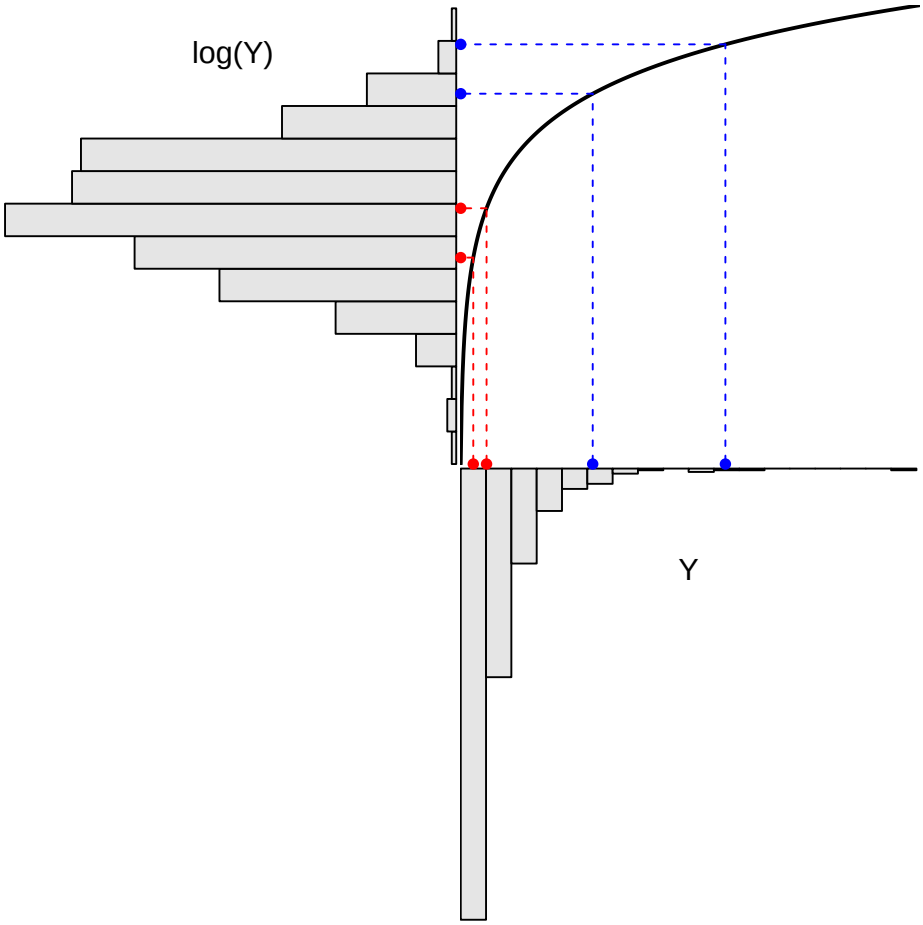


Figure 8.1: Demonstration of the result (upper-left) from applying the natural log transformation function (black curve in upper-right) to strongly right-skewed original values (lower-right).

Thus, the square-root transformation is not a "strong enough" transformation to normalize this strongly right-skewed original distribution.

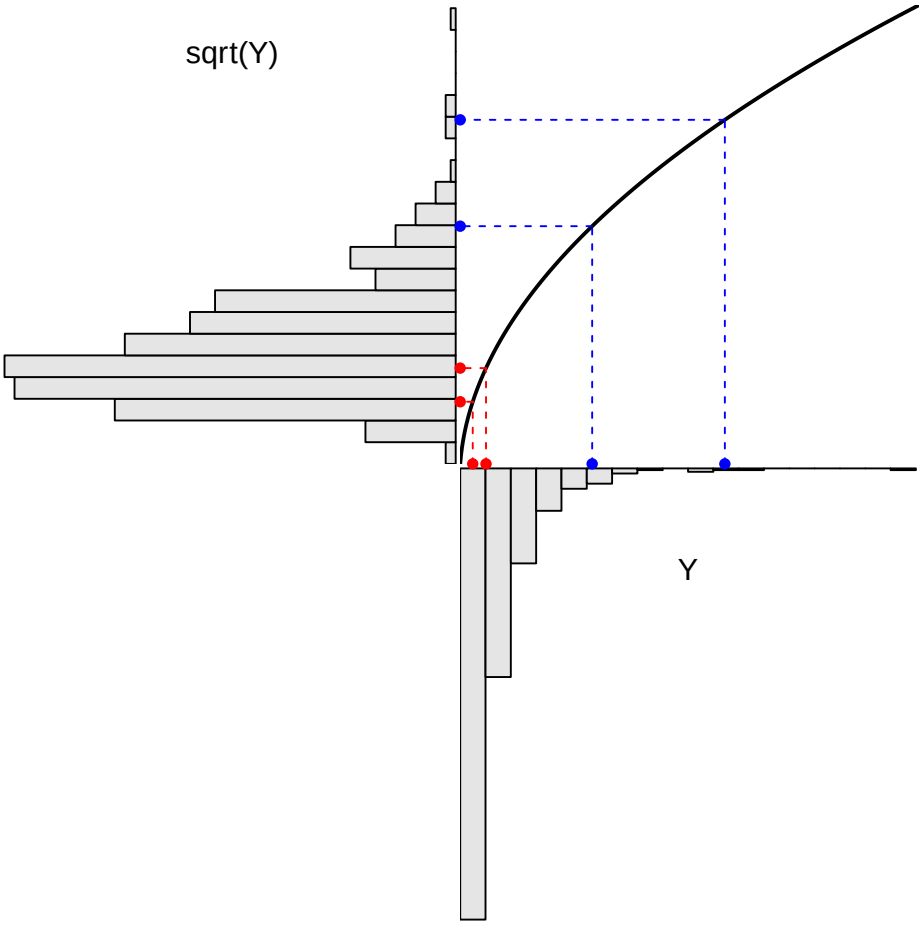


Figure 8.2: Demonstration of the result (upper-left) from applying the square root transformation function (black curve in upper-right) to strongly right-skewed original values (lower-right). The original values here are the same as those in the previous Figure.

The square root transformation is more likely to be useful when the original distribution is less strongly skewed (Figure 8.3).

As demonstrated above, the common transformations listed in Table 8.1 vary in their ability to "normalize" skewed distributions. Generally the common transformations in Table 8.1 are ordered from least to most powerful. In other words,

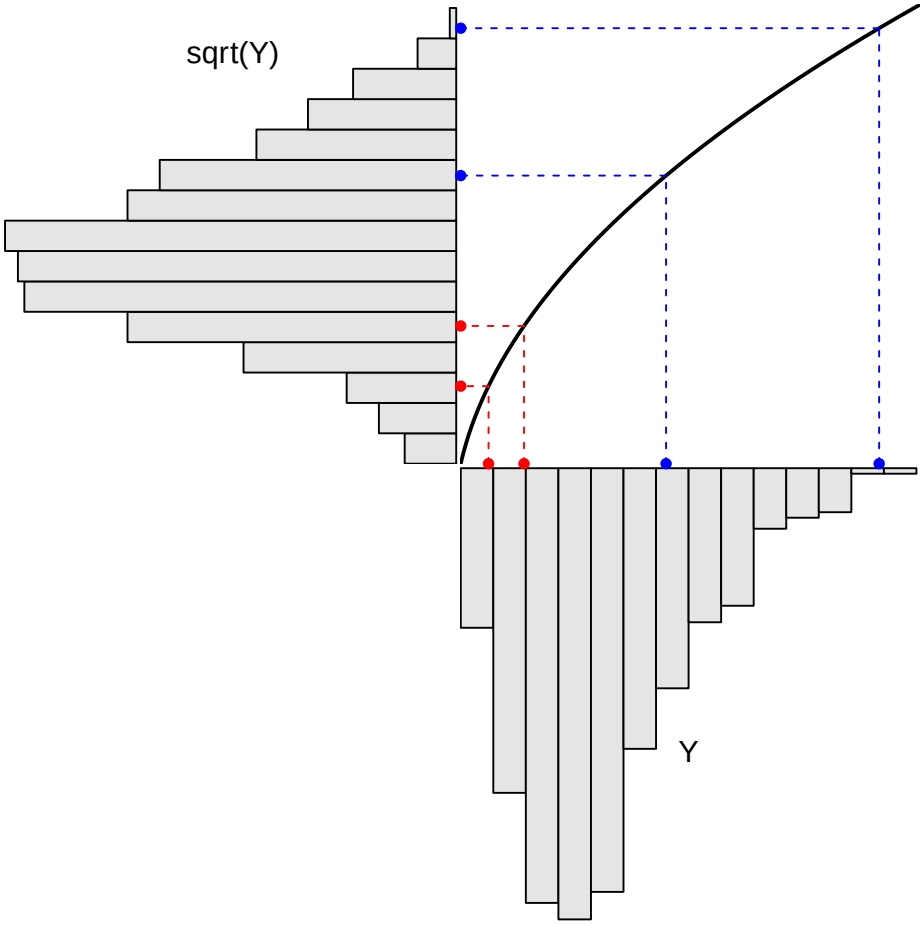


Figure 8.3: Demonstration of the result (upper-left) from applying the square root transformation function (black curve in upper-right) to slightly right-skewed original values (lower-right). The original values here are **not** the same as those in the previous two figures.

the transformations are ordered from those that "normalize" mildly skewed data to those that "normalize" strongly skewed data.[1]

It is possible to "combine" one of the common powers with the inverse transformation to create a larger array of transformations. For example, λ=-0.5 is an inverse square-root transformation. These types of transformations are common but less common than those listed in Table 8.1.

Table 8.2: Common inverse power transformations in ANOVAs. These transformations are much less common than those in Table 8.1.

Power	Name	Formula	R Code
λ=-0.25	Inverse Fourth Root	$Y^{-0.25} = \frac{1}{\sqrt[4]{Y}}$	`df$ifourrt.y <- df$Y^(-1/4)`
λ=-0.33	Inverse Cube Root	$Y^{-0.33} = \frac{1}{\sqrt[3]{Y}}$	`df$icubert.y <- df$Y^(-1/3)`
λ=-0.5	Inverse Square Root	$Y^{-0.5} = \frac{1}{\sqrt{Y}}$	`df$isqrt.y <- 1/sqrt(df$Y)`

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 all values become positive. This is called *shifting* the data and does not effect the shape of the distribution. In addition, power transformations are not effective if the range of values of the response variable is narrow.[2]

There are several methods to identify the power transformation that is most likely to meet the assumptions of a one-way ANOVA.[3] 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.

The trial-and-error method is made easier with software. For example, the `assumptionCheck()` function introduced in Section 7.5 can be used to quickly

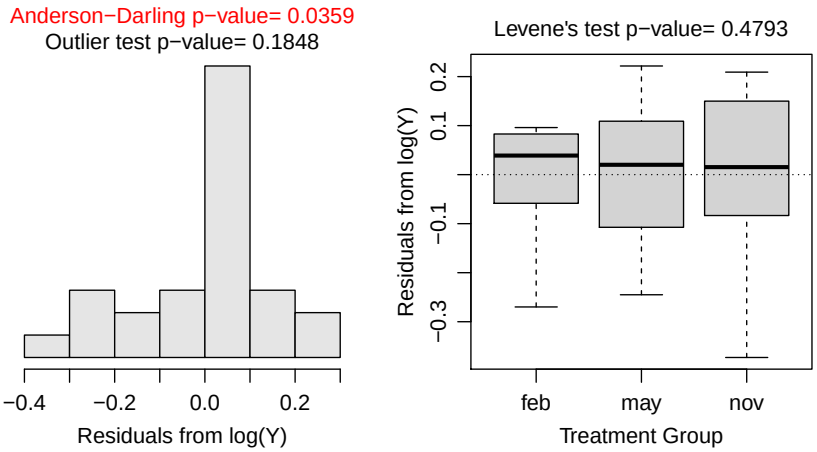
[1] Alternatively, the transformations are listed in order from the transformations that "spread out" the small values the least to those that "spread out" the small values the most.

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

[3] 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 SSwithin). A slightly modified Box and Cox approach is implemented in R by sending a `lm` object to `boxcox()` from the `MASS` package.

compute the graphs and hypothesis tests for assumption checking *after* the data have been transformed by the power given in the `lambday=` argument. For example, the code and results below show what the assumption checking would look like if the immunoglobulin measurements for the New Zealand opossums had been log transformed.

```
lm1 <- lm(imm~season,data=opp)
assumptionCheck(lm1,lambday=0)  #lambda=0 corresponds to log-transformation
```



Of course, it was shown in Section 7.5 that the assumptions for a One-Way ANOVA with these data had been met; thus, there is no need to explore a transformation here. However, this shows how easily one quickly test various transformations for a given set of data.

Note that `assumptionCheck()` only transform the data "behind-the-scenes." If you want to continue with transformed data then you need to use R code like that shown in the last columns of Tables 8.1 and 8.2.

8.2 Transformations from Theory

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 (i.e., $\sin^{-1}(\sqrt{Y})$). Also, a possible power transformation may be chosen from theory related to the response variable. For example, it is common to square-root transform response variables that are areas and cube-root transform response variables that are volumes. In addition, discrete counts are often transformed with the square root.

8.3 Interpretations After Transformations

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

First, tell the reader what transformation you used and how you arrived at it. In other words, clearly demonstrate that the assumptions were not met on the original scale and demonstrate that they were met on the transformed scale.

Second, when making a conclusion from the One-Way ANOVA p-value, refer to the transformed response variable in your conclusions. In other words, say "the mean square root of the response variable differed among groups" rather than "the mean of the response variable differed among groups." It will be implied that the means differed on the original scale, but you strictly tested on the transformed scale so you should be explicit about that here.

Third, back-transform estimates (and confidence intervals) for "points". In One-Way ANOVA this means that you should back-transform estimates of means. For example, if the response variable was log-transformed such that the **mean log of Y** was 1.356 then this should be back-transformed to the original scale with e1.356=3.881. Similarly if the response variable was square-root transformed such that the **mean square-root of Y** was 1.356 then this should be back-transformed to the original scale with 1.3562=1.839.

Fourth, do **NOT** back-transform "differences" unless those differences are from a log-transformation. In a One-Way ANOVA this translates into whether or not you can back-transform *differences* in means, as would result from multiple comparisons. For example if the result is a difference in the mean square-root of the response variable between groups then do not back-transform this value as it can **not** be back-transformed to anything meaningful. However, if the result is a difference in mean log of the response variable between groups then this **difference** can (and should) be back-transformed to something meaningful.

We know from algebra class that the difference in the log of two values is the log of the ratio of the two values – i.e., $log(a) - log(b) = log(\frac{a}{b})$. Thus, back-transforming the difference in log values results in a ratio of values on the original scale – i.e., $e^{log(a)-log(b)} = e^{log(\frac{a}{b})} = \frac{a}{b}$. Thus, in a One-Way ANOVA, the back-transformed **difference** in group means on the log scale is the **ratio** of group means on the original scale.

For example, suppose that the difference in log means for two groups is 0.5. Back-transforming this value gives $e0.5$=1.649. Thus, on the original scale, the mean for the first group is 1.649 times larger than the mean for the second group.

Alternatively, suppose that the difference in log means for two groups is -0.5.

Back-transforming this value gives e-0.5=0.607. Thus, on the original scale, the mean for the first group is 0.607 as large as the mean for the second group. Or, the mean for the *second* group is $\frac{1}{0.607}$=1.649 times larger than the mean for the *first* group.

Log-transformations are very special, as they allow both means and differences in means to be back-transformed to the original scale with a meaningful result. Because of this, the first transformation that you should try in all situations is the log-transformation. Many times we prefer the log-transformation over other transformations, even if the assumptions are not perfectly met with the log-transformation.

Always try the log-transformation first as it allows for meaningful back-transformations of means and differences.

If the log transformation does not work to meet the assumptions of the One-Way ANOVA then you should start with the least strong transformation (i.e., the square root) and successively move through more strong transformations until the assumptions are adequately met.

8.4 Back-Transformations in R

Transformations and back-transformations will be illustrated more in this module's assignment and in Module 9. However, the `emmeans()` function introduced in Section 6.3 makes back-transformations very easy.

While a transformation was not needed for the New Zealand opossums example, let's suppose for illustrative purposes that a square root transformation was used. First, a square root variable is created and a new ANOVA model with this variable is used.

```
opp$sqrtimm <- sqrt(opp$imm)
lm2 <- lm(sqrtimm~season,data=opp)
anova(lm2)
```

```
#R>   Analysis of Variance Table
#R>
#R>   Response: sqrtimm
#R>             Df   Sum Sq  Mean Sq F value    Pr(>F)
#R>   season     2 0.099654 0.049827  14.675 6.868e-05
#R>   Residuals 24 0.081487 0.003395
```

Multiple comparisons can be created as before, but there is an advantage to telling `emmeans()` that you used a square root transformation with `tran=` as shown below. Here we can see (see the two notes under the `$emmeans` and `$contrasts` portions of the output) that `summary()` reminds you that the results

are on the square root scale. This should help you remember to interpret these results on the square root scale.

```
mct <- emmeans(lm2,specs=pairwise~season,tran="sqrt")
( mcsumt <- summary(mct,infer=TRUE) )
```

```
#R>   $emmeans
#R>    season emmean     SE df lower.CL upper.CL t.ratio p.value
#R>    feb     0.816 0.0168 24    0.781    0.851 48.513  <.0001
#R>    may     0.849 0.0220 24    0.803    0.894 38.528  <.0001
#R>    nov     0.698 0.0206 24    0.656    0.741 33.886  <.0001
#R>
#R>   Results are given on the sqrt (not the response) scale.
#R>   Confidence level used: 0.95
#R>
#R>   $contrasts
#R>    contrast  estimate     SE df lower.CL upper.CL t.ratio p.value
#R>    feb - may  -0.0325 0.0277 24  -0.1017   0.0367 -1.172  0.4806
#R>    feb - nov   0.1179 0.0266 24   0.0515   0.1843  4.434  0.0005
#R>    may - nov   0.1504 0.0302 24   0.0751   0.2257  4.988  0.0001
#R>
#R>   Note: contrasts are still on the sqrt scale
#R>   Confidence level used: 0.95
#R>   Conf-level adjustment: tukey method for comparing a family of 3 estimates
#R>   P value adjustment: tukey method for comparing a family of 3 estimates
```

More importantly if you declare your transformation with `tran=` then you can use `summary()` to **appropriately** back-transform the results by including `type="response"`. Here, the `$emmeans` results are back-transformed as the *"Intervals are back-transformed from the sqrt scale"* note indicates, but you are reminded that the p-values were computed on the transformed scale with the *"Tests are performed on the sqrt scale"*. The `$contrasts` results are left on the square root scales as noted with *"Note: contrasts are still on the sqrt scale"* because `emmeans()` is smart enough to know that you should not back-transform *differences* when using a square root transformation.

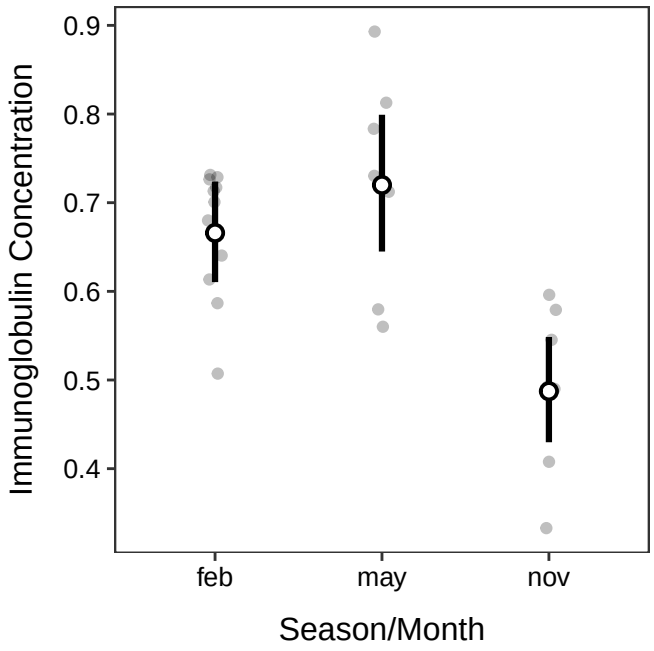
```
( mcsumbt <- summary(mct,infer=TRUE,type="response") )
```

```
#R>   $emmeans
#R>    season response     SE df lower.CL upper.CL t.ratio p.value
#R>    feb       0.666 0.0275 24    0.610    0.724 48.513  <.0001
#R>    may       0.720 0.0374 24    0.645    0.799 38.528  <.0001
#R>    nov       0.487 0.0288 24    0.430    0.549 33.886  <.0001
#R>
#R>   Confidence level used: 0.95
#R>   Intervals are back-transformed from the sqrt scale
```

```
#R>   Tests are performed on the sqrt scale
#R>
#R>   $contrasts
#R>    contrast  estimate     SE df lower.CL upper.CL t.ratio p.value
#R>    feb - may  -0.0325 0.0277 24  -0.1017   0.0367 -1.172  0.4806
#R>    feb - nov   0.1179 0.0266 24   0.0515   0.1843  4.434  0.0005
#R>    may - nov   0.1504 0.0302 24   0.0751   0.2257  4.988  0.0001
#R>
#R>   Note: contrasts are still on the sqrt scale
#R>   Confidence level used: 0.95
#R>   Conf-level adjustment: tukey method for comparing a family of 3 estimates
#R>   P value adjustment: tukey method for comparing a family of 3 estimates
```

The back-transformed means in `$emmeans` can be used to construct a plot of means as before, but you must realize that the means are now labeled as "response" (look at the last output above) rather than "emmean" (thus you must change `y=` in `geom_pointrange()`).

```
ggplot() +
  geom_jitter(data=opp,mapping=aes(x=season,y=imm),
              alpha=0.25,width=0.05) +
  geom_pointrange(data=mcsumbt$emmeans,
                  mapping=aes(x=season,y=response,ymin=lower.CL,ymax=upper.CL),
                size=1.1,fatten=2,pch=21,fill="white") +
  labs(y="Immunoglobulin Concentration",x="Season/Month") +
  theme_NCStats()
```



Now, look at the same results for a log transformation. Once again the results when not using `type="response"` show a note reminding you that the results are on the log scale.

```
opp$logimm <- log(opp$imm)
lm3 <- lm(logimm~season,data=opp)
mct <- emmeans(lm2,specs=pairwise~season,tran="log")
( mcsumt <- summary(mct,infer=TRUE) )
```

```
#R>   $emmeans
#R>    season emmean     SE df lower.CL upper.CL t.ratio p.value
#R>    feb     0.816 0.0168 24    0.781    0.851 48.513  <.0001
#R>    may     0.849 0.0220 24    0.803    0.894 38.528  <.0001
#R>    nov     0.698 0.0206 24    0.656    0.741 33.886  <.0001
#R>
#R>   Results are given on the log (not the response) scale.
#R>   Confidence level used: 0.95
#R>
#R>   $contrasts
#R>    contrast  estimate     SE df lower.CL upper.CL t.ratio p.value
#R>    feb - may  -0.0325 0.0277 24  -0.1017   0.0367 -1.172  0.4806
#R>    feb - nov   0.1179 0.0266 24   0.0515   0.1843  4.434  0.0005
```

```
#R>    may - nov   0.1504 0.0302 24   0.0751   0.2257  4.988  0.0001
#R>
#R>   Results are given on the log (not the response) scale.
#R>   Confidence level used: 0.95
#R>   Conf-level adjustment: tukey method for comparing a family of 3 estimates
#R>   P value adjustment: tukey method for comparing a family of 3 estimates
```

However, when `type="response"` is used, the `$emmeans` portion of the results are back-transformed with a note again saying so. However, the `$contrasts` portion is now also back-transformed because `emmeans()` is smart enough to know that it is possible (and useful) to back-transform differences from the log scale. In fact the rows in the `$contrasts` portion are appropriately labeled as **ratios** to aid your interpretation.

```
( mcsumbt <- summary(mct,infer=TRUE,type="response") )
```

```
#R>   $emmeans
#R>    season response     SE df lower.CL upper.CL t.ratio p.value
#R>    feb        2.26 0.0380 24     2.18     2.34 48.513  <.0001
#R>    may        2.34 0.0515 24     2.23     2.44 38.528  <.0001
#R>    nov        2.01 0.0414 24     1.93     2.10 33.886  <.0001
#R>
#R>   Confidence level used: 0.95
#R>   Intervals are back-transformed from the log scale
#R>   Tests are performed on the log scale
#R>
#R>   $contrasts
#R>    contrast  ratio     SE df lower.CL upper.CL t.ratio p.value
#R>    feb / may 0.968 0.0268 24    0.903     1.04 -1.172  0.4806
#R>    feb / nov 1.125 0.0299 24    1.053     1.20  4.434  0.0005
#R>    may / nov 1.162 0.0351 24    1.078     1.25  4.988  0.0001
#R>
#R>   Confidence level used: 0.95
#R>   Conf-level adjustment: tukey method for comparing a family of 3 estimates
#R>   Intervals are back-transformed from the log scale
#R>   P value adjustment: tukey method for comparing a family of 3 estimates
#R>   Tests are performed on the log scale
```

Chapter 9

One-Way Summary

Specific parts of a full One-Way ANOVA analysis have been described in Modules 5-8. In this module, a workflow for a full analysis is offered and that workflow is demonstrated with several examples.

9.1 Suggested Workflow

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 special attention to the distributional shape, center, dispersion, and outliers within each level of the grouping variable.
2. Show the sample size per group and comment on whether the study was balanced (i.e., same sample size per group) or not.
3. Address the independence assumption.
    - If this assumption is not met then other analysis methods must be used.
4. Fit the untransformed full model (i.e., separate group means) with `lm()`.
5. Check the other three assumptions for the untransformed model with `assumptionCheck()`.
    - Check equality of variances with a Levene's test and residual plot.
    - Check normality of residuals with an Anderson-Darling test and histogram of residuals.
    - Check for outliers with an outlier test, residual plot, and histogram of residuals.

6. If an assumption or assumptions are violated, then attempt to find a transformation where the assumptions are met.
   - Use the trial-and-error method with `assumptionCheck()`, theory, or experience to identify a possible transformation. Always try the log transformation first.
   - If only an outlier exists (i.e., there are equal variances and normal residuals) and no transformation corrects the outlier then *consider* removing the outlier from the data set.
   - Fit the ultimate full model with the transformed response or reduced data set.
7. Construct an ANOVA table for the full model with `anova()` and make a conclusion about the equality of means from the p-value.
8. If differences among group means exist, then use a multiple comparison technique with `emmeans()` and `summary()` to identify specific differences. Discuss specific differences **using confidence intervals**.
   - If the data were log-transformed then discuss specific ratios of means using back-transformed differences (use `tran=` in `emmeans()` and `type="response"` in `summary()` to computed the back-transformations).
9. Create a summary graphic of observations with group means **on the original scale** and 95% confidence intervals using `ggplot()` and results from `emmeans()`.
10. Write an overall conclusion of your findings.

9.2 Nematodes (*No Transformation*)

Root-Knot Nematodes (*Meloidogyne* spp.) 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.

A science fair student designed an experiment to determine the effect of increased densities of nematodes on the growth of tomato seedlings. 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

$$\begin{aligned} \mathrm{H}_0 &: \mu_0 = \mu_{1000} = \mu_{5000} = \mu_{10000} \\ \mathrm{H}_\mathrm{A} &: \text{At least one pair of means is different} \end{aligned}$$

where μ is the mean growth of the tomato seedlings and the subscripts identify densities of nematodes (see below).

Table 9.1: Number of pots at each nematode density treatment.

density	Freq
0	4
1000	4
5000	4
10000	4

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 much 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.

This study was "balanced" as the number of pots at each nematode density is the same (Table 9.1). The sample size, though, is quite small.

Independence appears to be largely met in this experiment. Each tomato plant was planted in a randomly selected individual pot that had been randomly assigned a number of nematodes. Thus, the growth of an individual plant should not effect the growth of another individual plant. I would be concerned if the pots were not randomly placed around the laboratory. For example, if all pots of a certain treatment were placed in one corner of the laboratory then conditions in that one corner may affect growth for those pots. There is no indication that this happened, so I will assume that it did not.

Note the detailed discussion of independence.

Variances among treatments appear to be approximately equal (Levene's p=0.8072; Figure 9.1-Right), the residuals appear to be approximately normally distributed (Anderson-Darling p=0.9268) and the histogram of residuals does not indicate any major skewness (Figure 9.1-Left), and there does not appear to be any major outliers in the data (outlier test p=0.6712; Figure 9.1). The analysis will proceed with untransformed data because the assumptions of the one-way ANOVA were met.

Table 9.2: ANOVA results for tomato seedling growth at four nematode densities.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
density	3	100.65	33.55	12.08	0.00062
Residuals	12	33.33	2.78		

Note the succinct writing with respect to testing the assumptions.

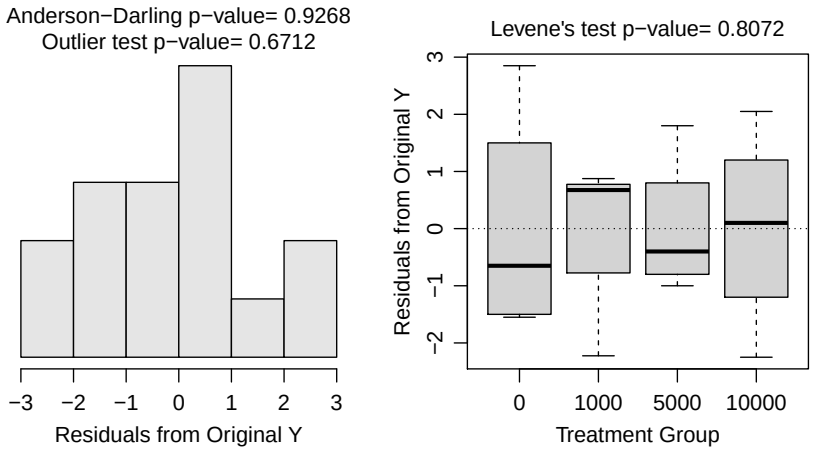


Figure 9.1: Histogram (Left) and boxplot (Right) of residuals from the utransformed One-Way ANOVA model for the tomato seedling growth at each nematode density.

There appears to be a significant difference in mean tomato seedling growth among at least some of the four treatments (p=0.0006; Table 9.2).

Note the use of ANOVA table to first identify any differences.

It appears that mean growth of tomatoes does not differ at densities 0 and 1000 (p=0.9974) and 5000 and 10000 (p=0.9992), but does differ for all other pairs of nematode densities (p 0.0070) (Table 9.3).

The student's hypothesis was generally supported (Figure 9.2). 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 increase from 1000 to 5000 nematodes per pot but not for increases from 0 to 1000 nematodes per pot or from 5000 to 10000 nematodes per pot. Specifically, it appears that the mean growth of the tomato seedlings is between 1.3 and 8.3 cm lower at a

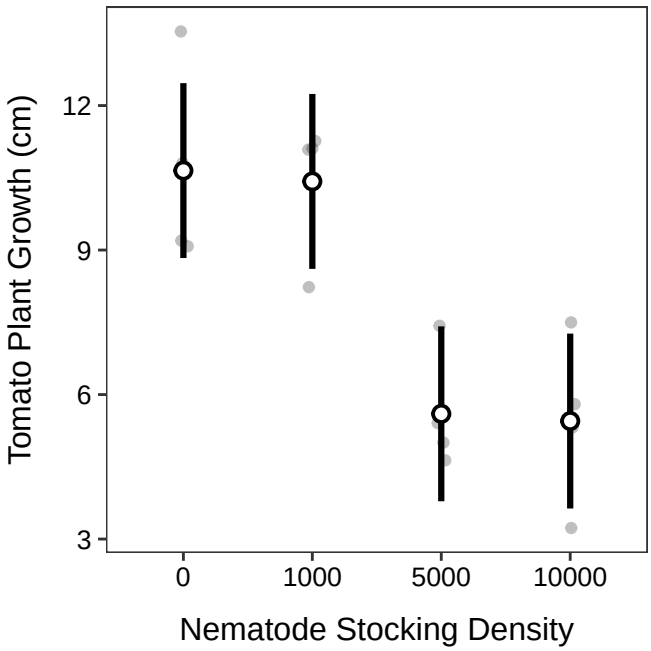
Table 9.3: Tukey's multiple comparisons for differences in mean tomato seedling growth for all pairs of four nematode densities.

contrast	estimate	lower.CL	upper.CL	p.value
0 - 1000	0.225	-3.274	3.724	0.9974
0 - 5000	5.050	1.551	8.549	0.0050
0 - 10000	5.200	1.701	8.699	0.0041
1000 - 5000	4.825	1.326	8.324	0.0070
1000 - 10000	4.975	1.476	8.474	0.0056
5000 - 10000	0.150	-3.349	3.649	0.9992

density of 5000 nematodes than at a density of 1000 nematodes (Table 9.2).

Note use of a confidence interval and specific direction (i.e., "lower") when describing the "difference."

\begin{figure}



{

}

\caption{Mean (with 95% confidence interval) of tomato growth at each nematode density.} \end{figure}

From this analysis, it appears that there is a "critical" density of nematodes

for tomato growth somewhere between 1000 and 5000 nematodes per pot. The experimenter may want to redo this experiment for densities between 1000 and 5000 nematodes per pot in an attempt to more specifically identify a "critical" nematode density below which there is very little affect on growth and above which there is a significant negative affect on growth.

R Code and Results

```
TN <- read.csv("http://derekogle.com/Book207/data/TomatoNematode.csv")
TN$density <- factor(TN$density)
lm.tn <- lm(growth~density,data=TN)
xtabs(~density,data=TN)
assumptionCheck(lm.tn)
```

```
anova(lm.tn)
mc.tn <- emmeans(lm.tn,specs=pairwise~density)
( mcsum.tn <- summary(mc.tn,infer=TRUE) )
ggplot() +
  geom_jitter(data=TN,mapping=aes(x=density,y=growth),
              alpha=0.25,width=0.05) +
  geom_pointrange(data=mcsum.tn$emmeans,
                  mapping=aes(x=density,y=emmean,ymin=lower.CL,ymax=upper.CL),
                size=1.1,fatten=2,pch=21,fill="white") +
  labs(x="Nematode Stocking Density",y="Tomato Plant Growth (cm)") +
  theme_NCStats()
```

9.3 Ant Foraging (*Transformation*)

Red wood ants (*Formica rufa*) forage for food (mainly insects and "honeydew" produced by aphids) both on the ground and in the canopies of trees. Rowan, Oak and Sycamore trees support very different communities of insect herbivores (including aphids) and it would be interesting to know whether the foraging efficiency of ant colonies is affected by the type of trees available to them. As part of an investigation of the foraging of *Formica rufa*, observations were made of the prey being carried by ants down trunks of trees. The total biomass of prey being transported was measured over a 30-minute sampling period on different tree specimens. The results were expressed as the biomass (dry weight in mg) of prey divided by the total number of ants leaving the tree to give a rate of food collection per ant per half hour. Observations were made on 28 Rowan, 26 Sycamore, and 27 Oak trees.[1]

[1]This example is directly from https://dzchilds.github.io/stats-for-bio/

Table 9.4: Number of ants recorded on each type of tree.

Tree	Freq
Oak	27
Rowan	28
Sycamore	26

The statistical hypotheses to be examined are

$$\mathrm{H}_0 : \mu_{Oak} = \mu_{Rowan} = \mu_{Sycamore}$$
$$\mathrm{H}_\mathrm{A} : \text{At least one pair of means is different}$$

where μ is the mean foraging rate of the ants and the subscripts identify the type of tree examined.

This study was slightly "unbalanced" as the number of ants recorded on each tree is similar, but not the same (Table 9.4).

The data appear to be independent as long as the separate trees were not in close proximity to each other. It is possible that the researchers examined one species of tree entirely within one "forest" of those trees. This would mean that the individuals within a species were somehow related, which would violate the independence assumption. In addition, the researchers could have looked at one specimen of each species all in a certain location, such that one tree could influence another tree. This again would violate the independence assumption. There is no indication that either of the scenarios occurred so I will assume that the individuals are independent both within- and among-groups.

Note careful discussion of independence here.

Variances among treatments for the untransformed data appear to be non-constant (Levene's p=0.0036; Figure 9.2-Right). The residuals appear to be not normally distributed (Anderson-Darling p=0.0002) with a fairly strong skew in the histogram (Figure 9.2-Left). There also appears to be a significant outlier (outlier test p=0.0012) that has a large residual and appears to be in the Oak group (Figure 9.2). None of the assumptions are met so transforming the food rate was considered.

A log transformation of foraging rate resulted in equal variances (Levene's p=0.4815; Figure 9.3-Right), normal (Anderson-Darling p=0.0603) or at least not skewed (Figure 9.3-Left) residuals, and no significant outliers (outlier test p>1). Thus, the assumptions of the One-Way ANOVA model appeared to have been adequately met on the log scale.

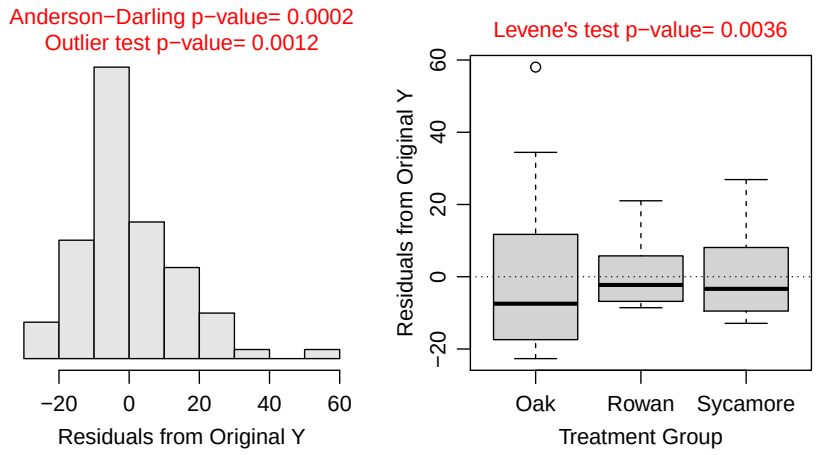


Figure 9.2: Histogram of residuals (Left) and boxplot of residual (Right) from the one-way ANOVA on untransformed foraging rate of ants.

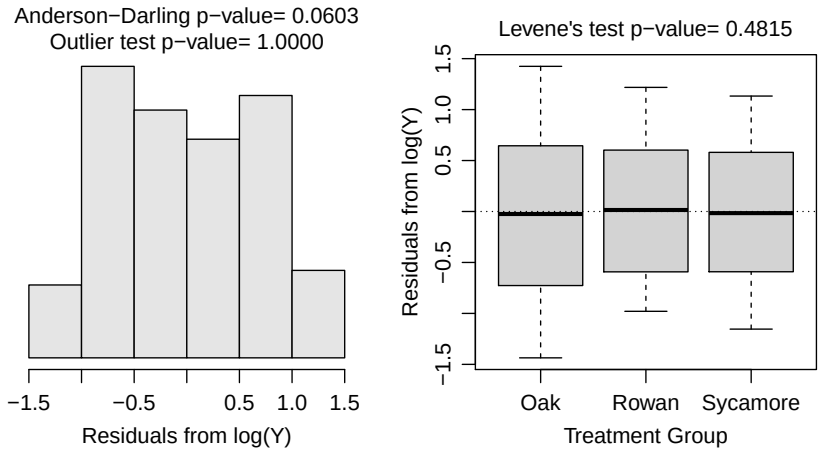


Figure 9.3: Histogram of residuals (Left) and boxplot of residual (Right) from the one-way ANOVA on log transformed foraging rate of ants.

Table 9.5: ANOVA results for the log transformed foraging rate of ants among tree species.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tree	2	7.479	3.739	7.287	0.00126
Residuals	78	40.028	0.513		

Table 9.6: Tukey's multiple comparisons for differences in mean log foraging rate of ants among tree species.

contrast	estimate	lower.CL	upper.CL	p.value
Oak - Rowan	0.738	0.276	1.199	0.0008
Oak - Sycamore	0.367	-0.103	0.837	0.1559
Rowan - Sycamore	-0.371	-0.837	0.096	0.1457

There appears to be a significant difference in mean log foraging rate of the ants among some of the tree species (p=0.0013; Table 9.5).

Note the careful use of the transformation name here.

Tukey multiple comparisons indicate that the mean log foraging rate was greater for ants on Oak trees than ants on Rowan trees (p=0.0008; Table 9.6). Specifically, the mean foraging rate for ants on an Oak tree was between 1.318 and 3.318 **times** higher than the mean foraging rate for ants on a Rowan tree (Table 9.7). However, there was no significant difference in mean log foraging rate between ants on Sycamore trees and ants on Oak (p=0.1559) or Rowan (p=0.1457; Table 9.6) trees.

Note how the back-transformed difference in means forms a ratio that is interpreted as a multiplicative change.

These results indicate that there is a difference in the mean foraging rate of ants, but only between those on Oak and Rowan trees, where ants on Oak trees had approximately double the mean foraging rate as ants on Rowan trees

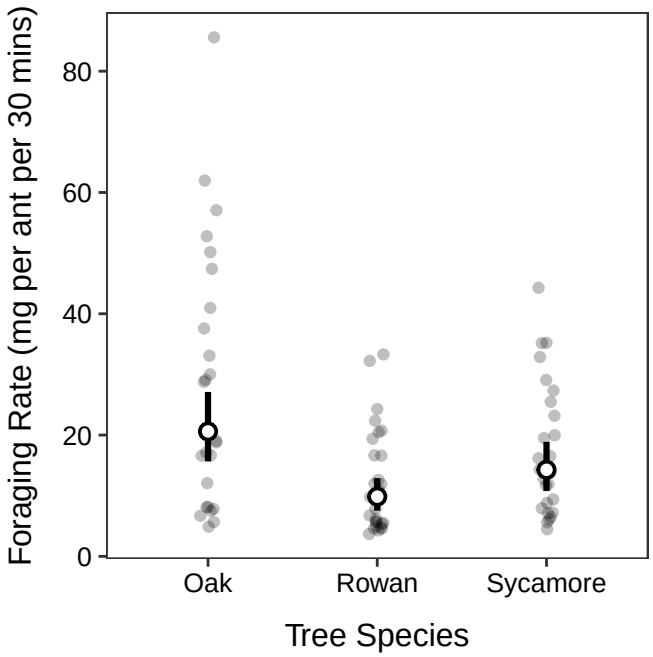
Table 9.7: Tukey's multiple comparisons for **ratios** of mean foraging rate of ants among tree species.

contrast	ratio	lower.CL	upper.CL	p.value
Oak / Rowan	2.091	1.318	3.318	0.0008
Oak / Sycamore	1.443	0.902	2.310	0.1559
Rowan / Sycamore	0.690	0.433	1.100	0.1457

(Figure 9.3). There was no difference in mean foraging rates for ants on Sycamore trees and either Oak or Rowan trees.

It is possible with a log to back-transform both means and differences in means.

\begin{figure}



{

}

\caption{**Back-transformed** mean (with 95% confidence interval) foraging rate of ants on each tree species. The mean foraging rate differed only between Oak and Rown trees.} \end{figure}

R Code and Results

```
AF <- read.csv("http://derekogle.com/Book207/data/Ants.csv")
lm.af <- lm(Food~Tree,data=AF)
xtabs(~Tree,data=AF)
assumptionCheck(lm.af)
```

```
assumptionCheck(lm.af,lambday=0)
```

```
AF$logFood <- log(AF$Food)
lm.aft <- lm(logFood~Tree,data=AF)
anova(lm.aft)
mc.aft <- emmeans(lm.aft,specs=pairwise~Tree,tran="log")
( mcsum.aft <- summary(mc.aft,infer=TRUE) )
( mcsum.afbt <- summary(mc.aft,infer=TRUE,type="response") )
ggplot() +
  geom_jitter(data=AF,mapping=aes(x=Tree,y=Food),
              alpha=0.25,width=0.05) +
  geom_pointrange(data=mcsum.afbt$emmeans,
                  mapping=aes(x=Tree,y=response,ymin=lower.CL,ymax=upper.CL),
                size=1.1,fatten=2,pch=21,fill="white") +
  labs(x="Tree Species",y="Foraging Rate (mg per ant per 30 mins)") +
  theme_NCStats()
```

9.4 Peak Discharge (*Transformation*)

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 was 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

$$\begin{aligned} \mathrm{H}_0 &: \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ \mathrm{H}_\mathrm{A} &: \text{At least one pair of means is different} \end{aligned}$$

where μ is the mean peak discharge estimate and the subscripts generically identify the four different methods for estimating peak discharge.

Each estimation method was used six times on the watershed and the resulting discharge estimates (in cubic feet per second) were recorded (Table 9.8).

At first glance, the data do not appear to be independent either within- or among-groups as the same methods were applied to the same watershed. However, the single watershed is the engineer's "population" of interest; thus, this form of data collection is not problematic unless the engineer (or you) attempt to make strict inferences to other watersheds. In addition, measurements from the same method may seem dependent, but this is the factor that is being examined, thus this is not a within-group dependency issue.

Table 9.8: Number of estimates for each method.

method	Freq
1	6
2	6
3	6
4	6

Note careful discussion of independence here.

Variances among estimation methods for the untransformed data appear to be non-constant (Levene's p=0.0136; Figure 9.4-Right). The residuals appear normally distributed (Anderson-Darling p=0.4106) or at least only slightly skewed (Figure 9.4-Left) and there are no significant outliers (outlier test p=0.3572). Unequal variances violates a critical assumption so transforming the discharge estimates was considered.

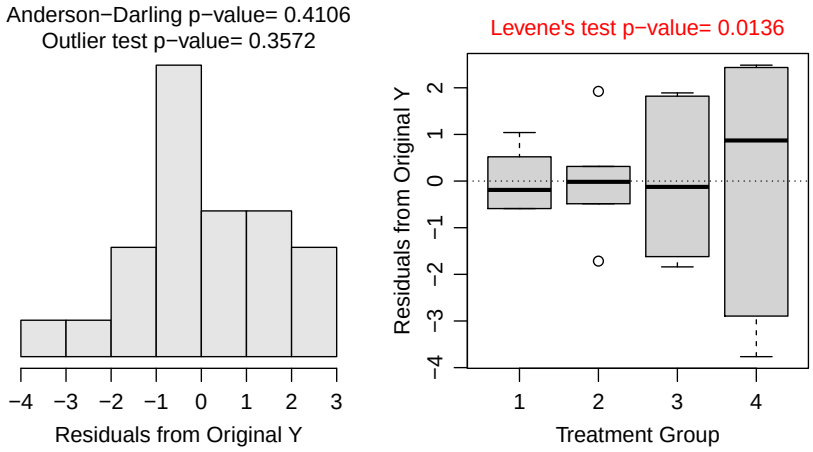


Figure 9.4: Histogram of residuals (Left) and boxplot of residual (Right) from the one-way ANOVA on untransformed peak discharge data.

A square root transformation for the peak discharge estimates resulted in equal variances (Levene's p=0.8680; Figure 9.5-Right), normal residuals (Anderson-Darling p=0.4207) or at least only slightly skewed (Figure 9.5-Left), and no significant outliers (outlier test p>1). Thus, the assumptions of the One-Way ANOVA model appear to have been adequately met on the square-root scale.

There appears to be a significant difference in mean square root peak

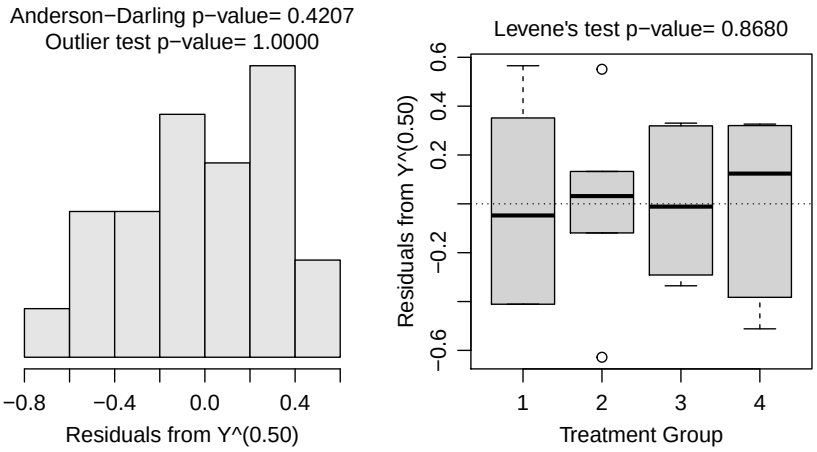


Figure 9.5: Histogram of residuals (Left) and boxplot of residual (Right) from the one-way ANOVA on square root transformed peak discharge data.

Table 9.9: ANOVA results for the square root peak discharge analysis by estimationg method.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
method	3	32.684	10.895	81.049	0
Residuals	20	2.688	0.134		

discharge among the four methods (p<0.00005; Table 9.9).

Note the careful use of the transformation name here.

Tukey multiple comparisons indicate that no two mean square root peak discharges were equal (p 0.0046; Table 9.10). It appears that the mean square root of estimated peak discharge increases from Method 1 to Method 2 to Method 3 to Method 4. For example, the mean square root of estimated peak discharge was between 2.471 and 3.656 units greater for Method 4 than for Method 1 (Table 9.10).

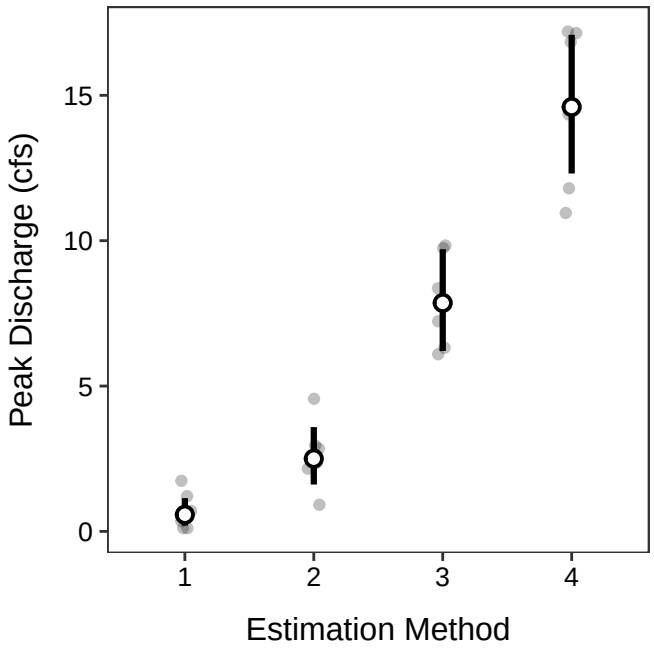
These results indicate that the mean peak discharge estimate differed significantly, with a higher peak discharge estimated for each method from Method 1 to Method 4 (Figure 9.4).

It is possible with a square root to back-transform means, but not differences in means.

\begin{figure}

Table 9.10: Tukey's multiple comparisons for differences in square root peak discharge for different estimation methods.

contrast	estimate	lower.CL	upper.CL	p.value
1 - 2	-0.825	-1.417	-0.232	0.0046
1 - 3	-2.046	-2.638	-1.453	0.0000
1 - 4	-3.063	-3.656	-2.471	0.0000
2 - 3	-1.221	-1.814	-0.629	0.0001
2 - 4	-2.239	-2.831	-1.646	0.0000
3 - 4	-1.018	-1.610	-0.425	0.0006



{

}

\caption{**Back-transformed** mean (with 95% confidence interval) estimates of peak discharge (cubic feet per second, cfs) by each estimation method. The transformed means were different among all estimationg methods.}
\end{figure}

R Code and Results

```
PD <- read.csv("http://derekogle.com/Book207/data/PeakDischarge.csv")
PD$method <- factor(PD$method)
```

```
lm.pd <- lm(discharge~method,data=PD)
xtabs(~method,data=PD)
assumptionCheck(lm.pd)
```

```
assumptionCheck(lm.pd,lambday=0.5)
```

```
PD$sqrtdischarge <- sqrt(PD$discharge)
lm.pdt <- lm(sqrtdischarge~method,data=PD)
anova(lm.pdt)
mc.pdt <- emmeans(lm.pdt,specs=pairwise~method,tran="sqrt")
( mcsum.pdt <- summary(mc.pdt,infer=TRUE) )
( mcsum.pdbt <- summary(mc.pdt,infer=TRUE,type="response") )
ggplot() +
  geom_jitter(data=PD,mapping=aes(x=method,y=discharge),
              alpha=0.25,width=0.05) +
  geom_pointrange(data=mcsum.pdbt$emmeans,
                  mapping=aes(x=method,y=response,ymin=lower.CL,ymax=upper.CL),
                size=1.1,fatten=2,pch=21,fill="white") +
  labs(x="Estimation Method",y="Peak Discharge (cfs)") +
  theme_NCStats()
```

TWO-WAY ANOVA

Chapter 10

Two-Way Conceptual Foundation

In contrast to a one-way ANOVA, a two-way ANOVA allows for simultaneously determining whether the mean of a quantitative response variable differs according to the levels of two grouping variables or an interaction between the two. For example, consider the following situations:

- Determining the effect of UV-B light intensity and tadpole density on mean growth of plains leopard frog (*Rana blairi*) tadpoles (Smith *et al.* 2000).
- Whether mean monoterpene levels in Douglas firs (*Pseudotsuga menziesii*) differ depending on exposure to ambient or elevated levels of CO2 and ambient or elevated levels of temperature (Snow *et al.* 2003).
- Whether mean microcystin-LR concentration (a common cyanotoxin) differed by time (days) and exposure type (control (no exposure), direct exposure, or indirect exposure) to duckweed (*Lemna gibba*) (LeBlanc *et al.* 2005).
- Determining if the mean Na-K-ATPase activity level differs among locations of the kidney and between normal and hypertensive rats (Garg *et al.* 1985).

The theory and application of two-way ANOVAs are discussed in this module. The presentation here depends heavily on the foundational material discussed for One-Way ANOVAs.

Table 10.1: Schematic of two OFAT experiments with UV-B light intensity and density of tadpoles. Tank numbers correspond to the random allocation of individual tanks to each treatment.

UVB	**Tanks**
High	37, 73, 63, 80, 76, 36, 67, 53, 83, 31, 29, 88, 62, 35, 21, 68, 17, 46, 89, 81, 26, 48
Low	85, 40, 30, 71, 60, 11, 47, 51, 24, 32, 54, 52, 23, 1, 38, 44, 78, 50, 3, 16, 9, 20

10.1 Two Factors

It is both possible and advantageous to manipulate more than one factor at a time to determine the effect of those factors on the response variable. For example, an experimenter may manipulate both the density of tadpoles and the level of exposure to UV-B light to determine the effect of both of these explanatory variables on body mass of tadpoles. The design of such an experiment and why it is beneficial to simultaneously manipulate two factors, as compared to varying each of those factors alone in two separate experiments, is examined in this section.

10.1.1 Design

Suppose that we are interested in the impact of two different UV-B light intensity levels (simply called "High" and "Low") and the density of tadpoles (1, 2, and 4 individuals per tank) on the body mass (g) of tadpoles. Further suppose that we have access to 90 tanks in which to conduct this work, where each tank will be stocked with a certain number of tadpoles and exposed to a certain UV-B light intensity. After a period of time the body mass of the tadpoles will be recorded (as a surrogate for growth).

Note here that in experiments like this the explanatory variables are often called **factors**. Thus, we are considering two factors in this study. Observational studies usually will use explanatory variable rather than factor.

One way to design this study would be to separate the 90 tanks available for the experiment into two sets. One set of tanks would be used to determine the effect of UV-B light intensity on tadpole body mass and the other set of tanks would be used to determine the effect of density on tadpole body mass. The tanks within each group would be randomly allocated to the different levels of each of the factors. This is called a "one factor at a time" (OFAT) design and might look like that shown in Table 10.1.

Density	**Tanks**
1 Tadpole	74, 57, 58, 7, 8, 86, 56, 43, 82, 90, 39, 27, 49, 77, 12
2 Tadpoles	61, 64, 15, 65, 6, 70, 10, 22, 14, 55, 72, 33, 25, 45, 42
4 Tadpoles	66, 59, 19, 28, 34, 18, 69, 4, 2, 79, 75, 5, 84, 87, 13

Table 10.2: Schematic depicting the crossing of three levels of UV-B light intensity with two levels of tadpole density to produce six treatments. Numbers in each cell represent individual tanks exposed to that treatment.

UVB	**Density**	**Tanks**
High	1 Tadpole	37, 73, 63, 80, 76, 36, 67, 53, 83, 31, 29, 88, 62, 35, 21
High	2 Tadpoles	68, 17, 46, 89, 81, 26, 48, 85, 40, 30, 71, 60, 11, 47, 51
High	4 Tadpoles	24, 32, 54, 52, 23, 1, 38, 44, 78, 50, 3, 16, 9, 20, 41
Low	1 Tadpole	74, 57, 58, 7, 8, 86, 56, 43, 82, 90, 39, 27, 49, 77, 12
Low	2 Tadpoles	61, 64, 15, 65, 6, 70, 10, 22, 14, 55, 72, 33, 25, 45, 42
Low	4 Tadpoles	66, 59, 19, 28, 34, 18, 69, 4, 2, 79, 75, 5, 84, 87, 13

In contrast, an alternative way to design this study would be to create "treatments"[1] where each level of one factor is combined with each level of the other factor. In this example, there would be six treatments as two levels of UV-B light intensity would be "crossed" with three levels of tadpole density. Tanks would be randomly allocated to treatments. This type of design is called a "completely crossed factorial design" (CCFD) and might look like that shown in Table 10.2.

Completely crossed factorial design (CCFD): A study design where each level of one explanatory variable (or factor) is combined with each level of the other explanatory variable (or factor) to form groups (or treatments).

10.2 Interaction Effects

An interaction among factors is said to exist if the effect of one explanatory variable on the mean response depends on the level of the other explanatory variable. For example, if the mean body mass of the tadpoles *increased* across densities in the low light intensity treatments but *decreased* across densities in the high light intensity treatments (Figure 10.1-Left), then an interaction between the density of tadpoles and UV-B light intensities is said to exist. However, if the pattern among tadpole densities is the same in both the low and high density light intensities (Figure 10.1-Right), then no interaction exists between the factors.

[1]The combination of levels is often called a "treatment" in an experiment. In an observational study it would just be called a "group."

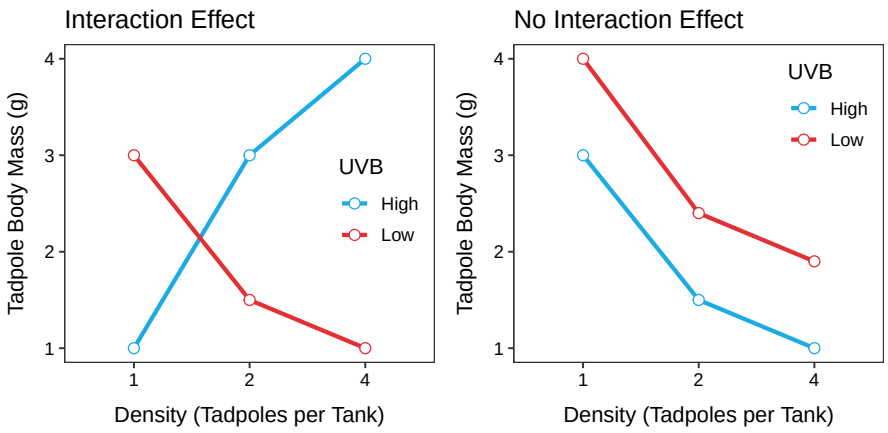


Figure 10.1: Interaction plot (mean growth rate for each treatment connected within the UV-B light intensity levels) for the tadpole experiment illustrating a hypothetical interaction effect (Left) and a lack of an interaction effect (Right).

**Interaction effect:: When the effect of one explanatory variable (or factor) on the mean response is significantly different among the levels of the other explanatory variable (or factor).

The tell-tale sign of an interaction effect is if you describe the effect of a factor on the mean response differently depending on the level of the other explanatory variable. For example, in Figure 10.2-Left the effect of tadpole density on mean body mass is increasing in the high UV-B light intensity treatments, but is neither increasing or decreasing in the low UV-B light intensity treatments. The fact that density had a different effect on mean body mass in the different UV-B light treatments tells of an interaction effect. In Figure 10.2-Right mean body mass increases with increasing density but at a very different rate (at least for densities of 1 to 2 tadpoles per tank). Again, this is a different conclusion about the effect of density on mean body mass depending on level of UV-B light; thus this represents an interaction effect.

In Module 11 we will develop objective mathematical measures of whether an interaction effect exists.[2] However, in addition to the concept above, an interaction effect will look like means that are connected for levels of one explanatory variable that "do not track together" for levels of the other explanatory variable. For example, in Figure 10.2-Left the "blue line" (connecting high UV-B light treatments) is steadily increasing whereas the "red line" (connecting low UV-B light treatments) is steadily flat – i.e., they

[2]That is, we will use a p-value.

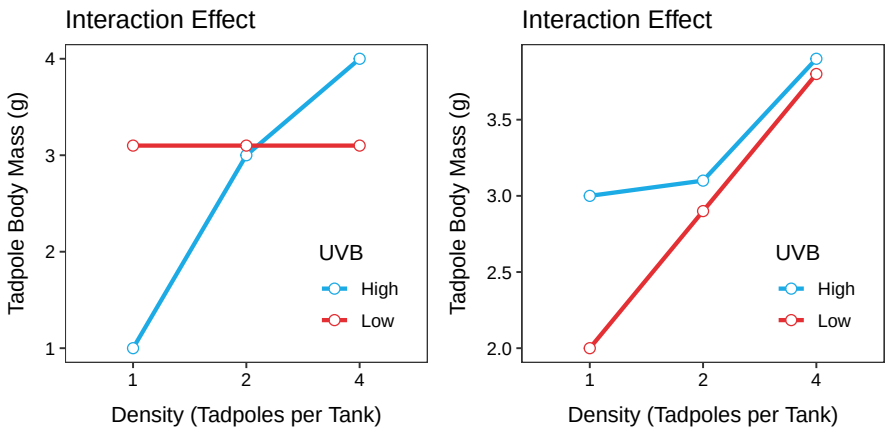


Figure 10.2: Interaction plot (mean growth rate for each treatment connected within the UV-B light intensity levels) for the tadpole experiment illustrating hypothetical interaction effects.

don't "track" together, which indicates an interaction effect. The two lines also do not track together in Figure 10.2 - Right.

An interaction effect is evident if the "lines" in an interaction plot do not largely "track together."

10.3 Main Effects

A main effect occurs when there is a difference in the mean response among levels of an explanatory variable *that is consistent across the levels of the other explanatory variable*. Thus, by definition, a main effect cannot exist if an interaction effect exists. Thus, do not try to identify main effects when an interaction is present.

Do **not** identify main effects when an interact effect exists.

For example, Figure 10.1-Right showed no significant interaction; thus, main effects can be assessed. In that case, the main effect of density is that mean body mass decreased when density increased from 1 to 2 tadpoles per tank and continued to decrease at a slower rate when the density increased from 2 to 4 tadpoles per tank. From the same figure, the effect of UV-B light density

is that mean body mass was greater at a low than at a high UV-B intensity.[3]

Main effects can also be determined for both interaction plots in Figure 10.3. The main effect for density in Figure 10.3-Left is the same as that described for Figure 10.1-Right. However, in this case there is no UV-B main effect because there is no separation in means by UV-B light level at each level of density. In other words the red line for low UV-B light and the blue line for high UV-B light are (nearly) directly on top of each other, so there is no difference in means due to UV-B light and thus no main effect of UV-B light.

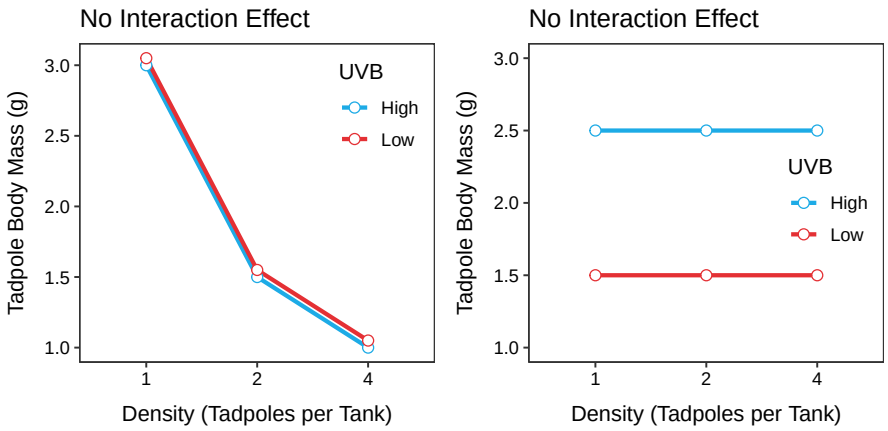


Figure 10.3: Interaction plot (mean growth rate for each treatment connected within the UV-B light intensity levels) for the tadpole experiment illustrating hypothetical lack2 of interaction effects.

In Figure 10.3-Right, there is no main effect of tadpole density as the mean body mass at each density within each UV-B light level is the same; i.e., mean body mass does not differ by density. However, there is a UV-B light main effect as mean body mass at the high UV-B light intensity is always[4] greater than mean body mass at the low UV-B light intensity.

As a reminder, only assess main effects if there is no interaction effect. Visually a main effect of the factor shown on the x-axis will be evident if the lines connecting the means are *not* horizontal (i.e., the means differ). In contrast, a main effect of the factor not shown on the x-axis (but shown as a legend) will be evident if the lines connecting the means are *not* right on top of each other.

[3] The blue line was always below the red line by roughly the same amount.
[4] For each density.

10.4 Advantages of CCFD

A two-factor CCFD has two major advantages over two separate OFAT experiments.

First, the CCFD allocates individuals more efficiently than the two OFAT experiments. For example, in the CCFD of the tadpole experiment, all 90 individuals were used to identify differences among the light intensities **AND** all 90 individuals were used to identify differences among the densities of tadpoles. However, in the OFAT experiments, only 44 individuals[5] were used to identify differences among the light intensities and only 45 individuals were used to identify differences among the tadpole densities.

CCFD studies effectively increase the sample size for identifying effects on the response variable.

Efficiency in allocation of individuals effectively leads to an increased sample size for determining the effect of any one factor. An increased sample size means more precise estimates of means[6] and more statistical power. Statistical power is the probability of correctly rejecting a false null hypothesis. In lay terms, statistical power is the ability to correctly identify a difference in means when a difference in means really exists. Relatedly, the increased sample size means that smaller effect sizes can be identified. In practice this means that smaller differences in means will be found to be statistically different. In yet other words, a smaller "signal" can be detected. Finally, the increased efficiency means that the same power and detectable effect size can be obtained with a smaller overall sample size and, thus, lower costs.

More efficient use of individuals in CCFD studies results in increased power and increased ability to detect small effect sizes, or lower costs to obtain the same power and effect size detectability.

A second advantage of CCFD studies is that they allow researchers to identify if an interaction effect exists between the two explanatory variables. An interaction effect cannot be detected in OFAT studies because the two explanatory variables are not considered simultaneously.

[5] One individual was not used in order to have equal numbers of individuals in each treatment.

[6] That is, a reduced SE

Chapter 11

Two-Way Analytical Foundation

In this module we will examine the analytical foundation of a Two-Way ANOVA. In many ways, this foundation is very similar to that for the One-Way ANOVA. However, there are some striking differences. The similarities and differences are discussed here.

11.1 Terminology

11.1.1 Definitions

Some terminology must be developed before discussing the objective criteria for determining the significance of main and interaction effects in a Two-Way ANOVA. The tadpole body mass study discussed in Module 10 was an experiment with two *factors*: UV-B light intensity with two *levels* (High and Low) and the density of tadpoles with three *levels* (1, 2, and 4 tadpoles). The combination of these two factors created six *treatments* with 15 individuals or *replicates* per treatment.[1]

11.1.2 Graphing

Two-factor studies can be visualized with the response variable on the y-axis, levels of one factor on the x-axis, and levels of the other factor shown with different colors or symbols. It does not make a difference which factor is on the x-axis, though putting the factor with more levels on the x-axis makes for a less cluttered graph.

[1] These terms are not defined specifically here as it is assumed that you were introduced to basic experimental design terms in your introductory statistics course.

In this module, two graphs will often be shown side-by-side. These are the same data but the role of the factor variables in the plot are reversed. For example, Figure 11.1-Left has tadpole density on the x-axis with different colors for the UV-B light levels and Figure 11.1-Right has UV-B light intensity on the x-axis with different colors for the tadpole densities.

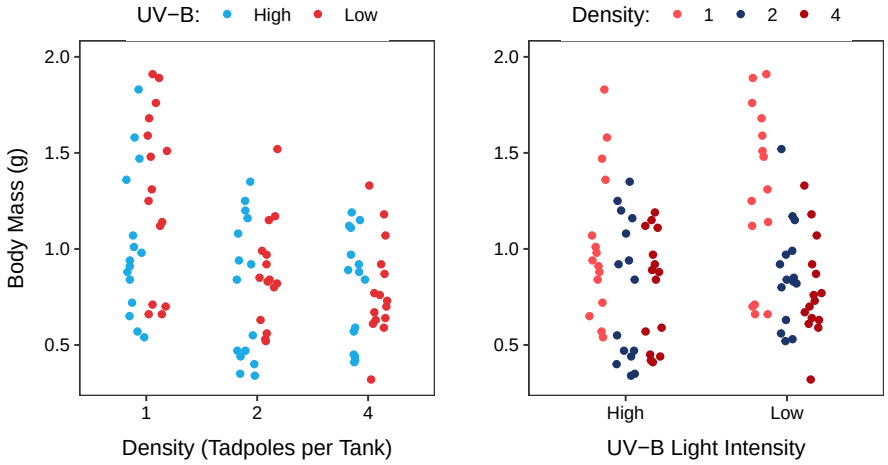


Figure 11.1: Tadpole body mass by density and UV-B light intensity factors. The points are jittered with respect to the x-axis so that each point can be seen. The two panels differ only in which factor variable is displayed on the x-axis.

11.1.3 Symbols

In a two-factor study, one of the factors is generically labeled as Factor A and the other factor is generically labeled as Factor B. Ultimately it does not make a difference which factor is considered first. In this example, tadpole density will be the first (A) factor and UV-B light intensity will be the second (B) factor.

Factor A has a levels and Factor B has b levels. In the tadpole example, a=3 (1, 2, and 4 tadpoles per tank) and b=2 (High and Low UV-B light intensity). We use i to index the first (A) factor and j to index the second (B) factor. Thus, i varies from 1 (first level) to a (last level) and j varies from 1 to b.

For the sake of simplicity the same number of replicates are used in each treatment.[2] With this m is the number of replicates *per treatment*. In the tadpole example m=15. We use k as an index for individuals **within** a treatment. Thus, k varies from 1 to m for each treatment.

[2] An experiment where each treatment has the same number of replicates is called a balanced design.

With these definitions, the response variable recorded on the kth individual in the treatment defined by the ith level of Factor A and the jth level of the Factor B is denoted by Y_{ijk}. For example, the body mass of tadpoles in the seventh tank (k=7) that received 1 tadpole (i=1) and low UV-B light (j=2) would be Y_{127} (Figure 11.2).

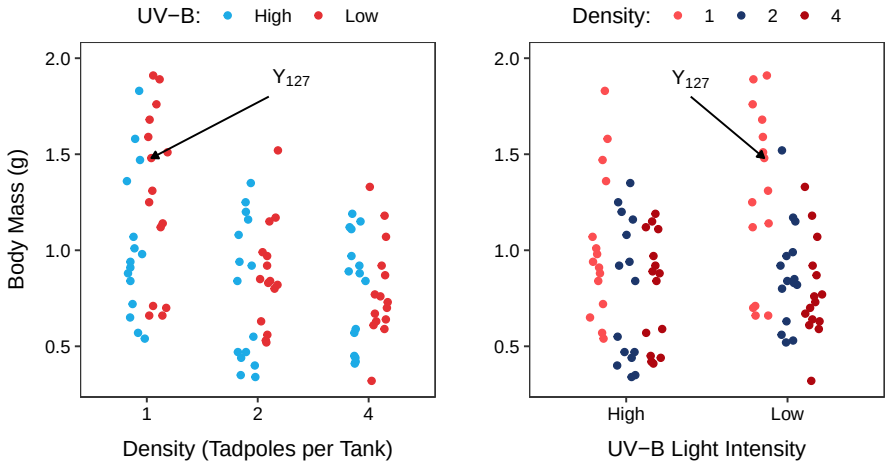


Figure 11.2: Tadpole body mass by density and UV-B light intensity factors. The points are jittered with respect to the x-axis so that each point can be seen. The two panels differ only in which factor variable is displayed on the x-axis.

The mean response for the ith level of Factor A and the jth level of Factor B is denoted by $\overline{Y}_{ij\cdot}$. The $\overline{Y}_{ij\cdot}$ are called **treatment means** in an experiment or **group means** in an observational study. The "dot" in $\overline{Y}_{ij\cdot}$ replaces the subscript in Y_{ijk} that was summed across when computing the mean. Treatment means are calculated by summing across individuals in a treatment (the k subscript), thus the k subscript is replaced with a dot. As an example, the treatment mean body mass for tadpoles in the 1 tadpole density (i=1) and low UV-B light intensity (k=2) would be $\overline{Y}_{12\cdot}$ (Figure 11.3).[3]

The mean response for the ith level of Factor A is given by $\overline{Y}_{i\cdot\cdot}$. These **level means** are calculated by first summing individuals in each treatment (the k subscript) and then summing across UV-B light levels (the j subscript); thus, both the k and j subscripts are replaced with a "dot." For example, the mean body mass for tadpoles in the 1 tadpole density (i=1) is $\overline{Y}_{1\cdot\cdot}$ (Figure 11.4-Left).

Similarly, the mean response for the jth level of Factor B is given by $\overline{Y}_{\cdot j\cdot}$. These **level means** are calculated by first summing individuals in each treatment (the k subscript) and then summing across tadpole density levels

[3]This figure is similar to the interaction plots from Module **??**.

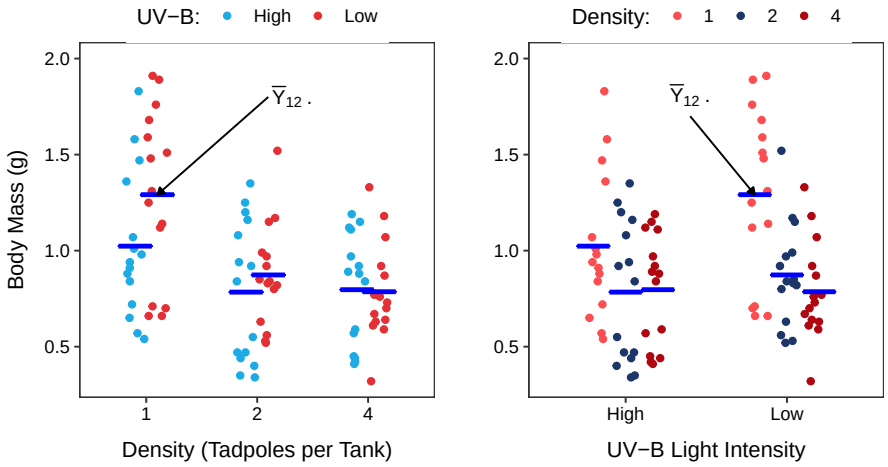


Figure 11.3: Same as previous figure except that six treatments means are shown with horizontal blue segments.

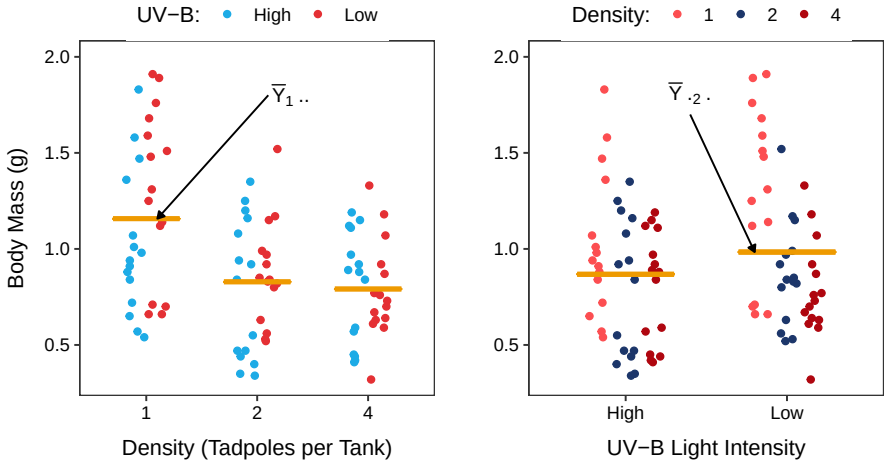


Figure 11.4: Same as previous figure except that three density level means are shown on the left and two UV-B light level means are shown on the right with horizontal orange segments.

(the i subscript); thus, both the k and i subscripts are replaced with a "dot." For example, the mean body mass for tadpoles in the low UV-B light intensity (j=2) is $\overline{Y}_{\cdot 2 \cdot}$ (Figure 11.4-Right).

Finally, the mean response regardless of level of any factor is given by $\overline{Y}_{\cdots}$ and is called the **grand mean** (Figure 11.5). The grand mean is calculated by summing individuals in each treatment, then summing across UV-B light intensities, and then summing across tadpole densities; thus, all three subscripts are replaced with a "dot."

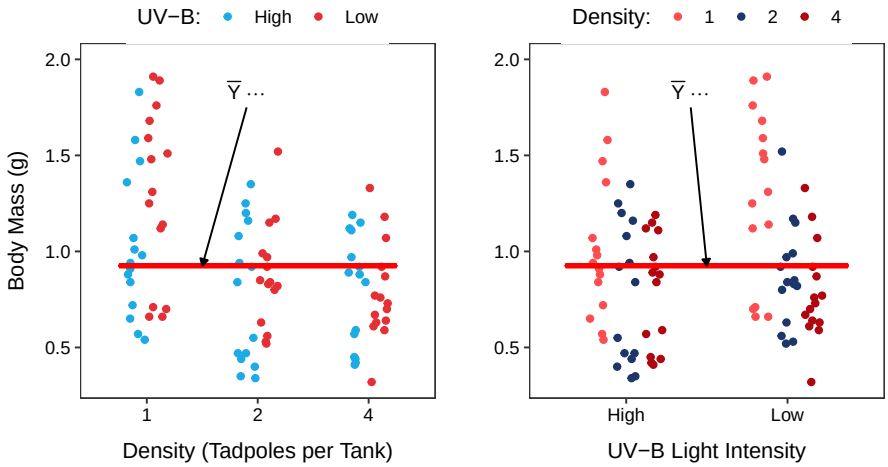


Figure 11.5: Same as previous figure except that the grand means is shown with a horizontal red segment in each panel.

In symbols for means, dots replace lettered subscripts for subscripts summed across when calculating the mean.

11.2 Total and Within SS, df, and MS

11.2.1 Models

The total and within SS and df[4] are effectively the same with a Two-Way ANOVA as with a One-Way ANOVA, though their calculation may look more complicated. In a Two-Way ANOVA, the *full model* uses a separate mean for each treatment group and the *simple model* uses a single grand mean for all

[4]Thus, also the MS.

treatment groups. In essence, the simple model says that each treatment mean should be modeled by a common mean (Figure 11.6-Left), whereas the full model says that each treatment mean should be modeled by a separate mean (Figure 11.6-Right). It should be evident that this is the same simple and full models used in a One-Way ANOVA.

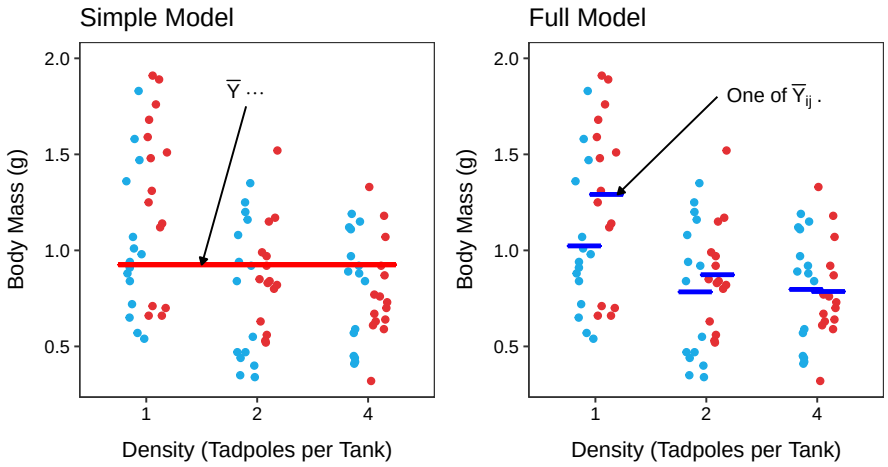


Figure 11.6: Tadpole body mass by density and UV-B light levels (different colored points) with the grand mean of the simple model (Left) and the treatments means of the full model (Right) shown.

11.2.2 SSTotal, dfTotal, and MSTotal

As discussed with a One-Way ANOVA, SSTotal measures the lack-of-fit of the observations around the simple model of a grand mean. Visually, this computation sums the square of the vertical distance of each point from the red line at the grand mean in Figure 11.6-Left; i.e.,

$$\text{SS}_{\text{Total}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(Y_{ijk} - \overline{Y}_{\cdots}\right)^2$$

This formula may appear intimidating but focus on the part being summed – $\left(Y_{ijk} - \overline{Y}_{\cdots}\right)^2$. This is simply the square of each observation (Y_{ijk}) from the simple model of a grand mean ($\overline{Y}_{\cdots}$). The three summations simply mean[5] to

[5] Read summations from right-to-left … in this case the summation across k and then across j and then across i

sum across individuals, then across levels of Factor B, and then across levels of Factor A. In other words, sum the squared residuals across *all* individuals, exactly what you did for a One-Way ANOVA.

The total degrees-of-freedom is still the total number of individuals (n) minus 1 because only one parameter (the grand mean) is being used in the simple model. Note, however, that $n = abm$, or the number of treatments (ab) times the number of replicates (or individuals) per treatment. Thus dfTotal=$abm - 1$.

The MSTotal is (as always) SSTotal divided by dfTotal and represents the variance of **individuals** around the grand mean (or simple model).

11.2.3 SSWithin, dfWithin, and MSWithin

Not surprisingly, SSWithin measures the lack-of-fit of the observations around the full model of separate means for each treatment. Visually, this computation sums the square of the vertical distance of each point from the blue line at the corresponding treatment mean in Figure 11.6-Right; i.e.,

$$\text{SS}_{\text{Within}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(Y_{ijk} - \overline{Y}_{ij\cdot}\right)^2$$

Again focus on the part of the formula being summed – $\left(Y_{ijk} - \overline{Y}_{ij\cdot}\right)^2$. This is the square of each observation (Y_{ijk}) from each treatment mean ($\overline{Y}_{ij\cdot}$). The three summations are the same as for SSTotal; i.e., sum the squared residuals across **all* individuals.

The within degrees-of-freedom is still the total number of individuals (n) minus the number of groups, because a separate mean is used for each treatment/group in the full model. Thus dfWithin=$abm - ab$ or dfWithin=$ab(m-1)$.

The MSWithin is (as always) SSWithin divided by dfWithin and represents the variance of **individuals** around the treatment/group means (or full model).

11.3 Among SS, df, and MS

The SSAmong is usually found by subtraction (i.e., SSTotal-SSWithin), which again indicates that SSAmong is the improvement in fit[6] between the full and simple models. It can also be shown that

[6]Really, the reduction in lack-of-fit

$$\text{SS}_{\text{Among}} = m \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{\cdots} \right)^2$$

Again focus on the part being summed, which is the square of the difference between each treatment mean ($\overline{Y}_{ij\cdot}$) and the grand mean ($\overline{Y}_{\cdots}$; Figure 11.7). Thus, as before, SSAmong measures how different the treatment means are. The rest of the formula simply sums the differences in means across all treatments[7] and then multiplies by m to account for the m individuals that went into calculating each treatment mean.[8]

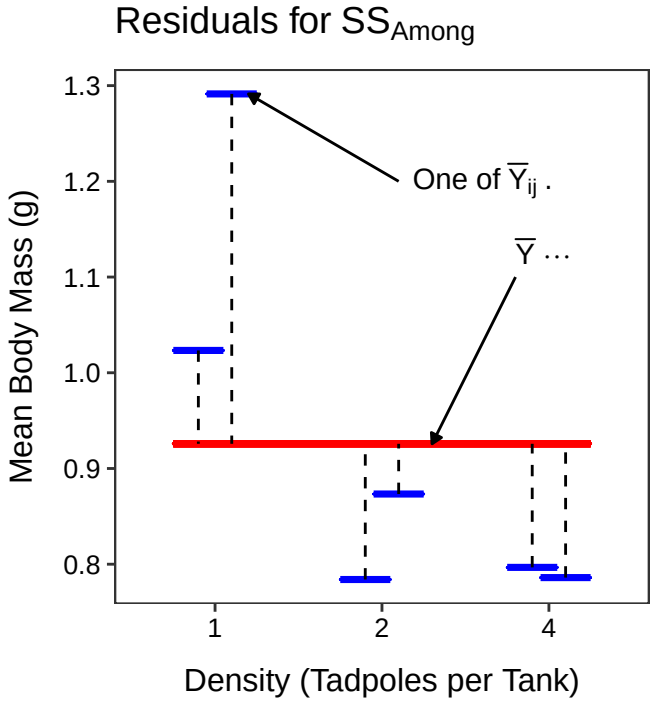


Figure 11.7: Mean tadpole body mass by density and UV-B light levels (not differentiated) with the grand mean of the simple model (red horizontal line) and the treatments means of the full model (blue horizontal lines) shown. Vertical dashed lines are "residuals" between the two types of means. Note that the y-axis scale is different than all previous plots.

[7] First across levels of Factor B and then across levels of Factor A.

[8] Multiplying by m scales the summation to the same number of individuals as summed in SSTotal and SSWithin; i.e., allowing a comparison of apples to apples.

The among df may be obtained by subtraction (i.e., dfAmong=dfTotal-dfWithin), which indicates that dfAmong measures the difference in complexity[9] between the simple and full models. It can also easily be shown that dfAmong=$ab-1$, or the number of treatments/groups minus 1, which is exactly as it was with the One-Way ANOVA.

Finally, MSAmong is equal to SSAmong divided by dfAmong and represents the variance of **treatment/group means**. Thus, the larger MSAmong is the more different the treatment/group means are. Of course, an F-ratio test statistic and corresponding p-value should be calculated to determine if MSAmong is "large" relative to MSWithin and whether we should conclude that there is a significant difference in treatment/group means.

11.3.1 Partitioning SSAmong

To this point, everything in a Two-Way ANOVA has been the same as it was for a One-Way ANOVA, just with a few more symbols. However, an issue occurs if H0 is rejected in favor of HA. If this occurs, then we would conclude that there is a significant difference in treatment/group means. However, as discussed in Module 10, a difference in treatment means could be related to differences in Factor A level means, differences in Factor B level means, or differences in means due to the interaction of Factor A and Factor B. Which factor, factors, or their interaction is responsible for the difference in treatments means must be teased out in an objective way.

Just as SSTotal partitioned into parts (i.e., SSAmong and SSWithin), SSAmong partitions into parts due to differences in the levels of Factor A, differences in the levels of Factor B, and differences in the interaction among the two factors. In other words,

$$SS_{Among} = SS_A + SS_B + SS_{A:B}$$

where A:B represents the interaction between Factor A and Factor B.

It can be shown that

$$\mathrm{SS_A} = mb\sum_{i=1}^{a}\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)^2$$

Again, focus on the part that is summed, which is the square of the differences in the Factor A level means and the grand mean (Figure 11.8-Left). Thus, SSA measures how different the Factor A levels means are, just as you would expect. Note that the rest of the formula says that you must sum across the

[9] Difference in number of estimated parameters in the models.

Factor A levels and then multiply by the number of individuals that went into calculating the Factor A levels means (i.e., m individuals across b levels of Factor B).

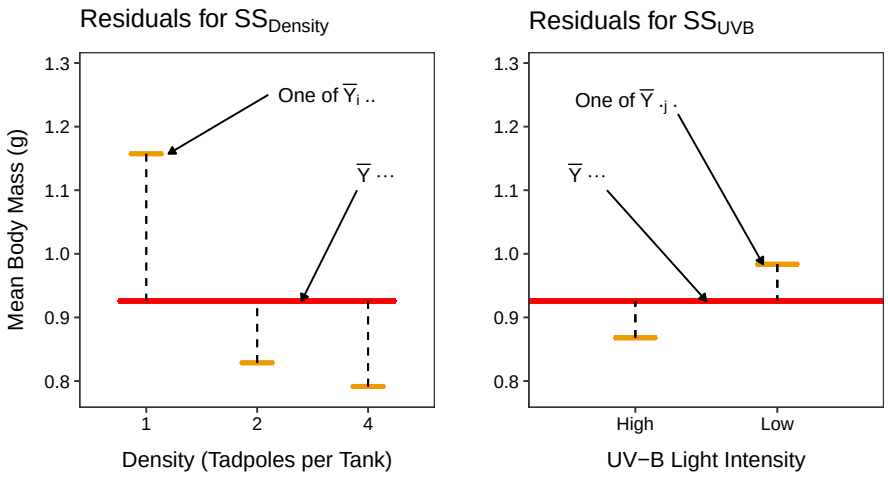


Figure 11.8: Mean tadpole body mass by density and UV-B light levels (not differentiated) with the grand mean of the simple model (red horizontal line) and the level means shown for the tadpole densities (Left) or UV-B light intensities (Right). Vertical dashed lines are "residuals" between respective level means and the grand means.

Similarly (Figure 11.8-Left),

$$\text{SS}_\text{B} = ma \sum_{j=1}^{b} \left(\overline{Y}_{\cdot j \cdot} - \overline{Y}_{\cdots} \right)^2$$

The interaction SS is difficult to describe or to visualize, but it is easily calculated by subtraction:

$$SS_{A:B} = SS_{Among} - SS_A - SS_B$$

The dfAmong partitions in the same way that SSAmong partitions; i.e., dfAmong=dfA+dfB+dfA:B. Further, dfA=$a-1$ and dfB=$b-1$; their respective number of level means minus 1, as you would expect. The dfA:B is most easily found by subtraction (dfA:B=dfAmong-dfA-dfB), but is also dfA:B=$(a-1)(b-1)$.)

Table 11.1: An ANOVA table for testing if mean body mass of tadpoles differs by density, UV-B light intensity, or the interaction between density and UV-B light. Note that the "Total" row is not shown.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
density	2	2.434	1.217	10.381	0.0001
uvb	1	0.300	0.300	2.563	0.1131
density:uvb	2	0.299	0.149	1.275	0.2847
Residuals	84	9.846	0.117		

Of course, MSA, MSB, and MSA:B are all computed by dividing the corresponding SS by the df. Thus, MSA is the variance explained by Factor A, or the difference in the Factor A level means. If MSA is "large" relative to MSWithin then there is likely a difference in the Factor A level means and there is a so-called Factor A main effect. The same argument can be made for Factor B.

The MSA:B is more difficult to describe, but can be thought of as the variance explained by the interaction between Factor A and Factor B. If MSA:B is "large" relative to MSWithin then there is likely a difference in means due to an interaction between factors A an B and there is a so-called interaction effect.

11.4 ANOVA Table

The F-ratio test statistics and corresponding p-values for the Factor A and Factor B main effects and the interaction between the two are summarized in an ANOVA table.

The following observations or conclusions can be drawn from Table 11.1.

- There are three levels of density (i.e., one more that dfdensity).
- There are two levels of UV-B light intensity (i.e., one more that dfuvb).
- There are 90 individuals (or replicates) (i.e., one more than dfTotal=2+1+2+84=89.)
- The variance of individuals around the grand mean is MSTotal=0.145 ($=\frac{2.434+0.300+0.299+9.846}{2+1+2+84} = \frac{12.879}{89}$).
- The variance of individuals around the treatment means is MSWithin=0.117.
- The variance of treatment means around the grand mean is MSAmong=0.607 ($=\frac{2.434+0.300+0.299}{2+1+2} = \frac{3.033}{5}$).
- The variance of density level means around the grand mean is MSdensity=1.217.
- The variance of UV-B level means around the grand mean is MSuvb=0.300.

- There is not a significant interaction effect (p=0.2847).
- There is not a significant UV-B light intensity main effect (p=0.1131).
- There is a significant tadpole density effect (p=0.0001), which means that the mean body mass differs for at least one pair of tadpole densities.[10]

[10] We will discuss methods to identify which pairs differ in Module 12.

Chapter 12

Two-Way Analysis

In this module, a thorough Two-Way ANOVA will be performed using the experiment introduced in Module 11, where the effect of tadpole density and UV-B light intensity on tadpole body mass was examined.

12.1 Model Fitting in R

The data are loaded into R below. Because `density` was recorded as a number (i.e., 1, 2, and 4) rather than as an obvious grouping (i.e., "one", "two", and "four"), it must be explicitly converted to a factor before it can be used in a Two-Way ANOVA Model.

```
tad <- read.csv("http://derekogle.com/Book207/data/Tadpoles.csv")
tad$density <- factor(tad$density)
str(tad)
```

```
#R>   'data.frame': 90 obs. of  3 variables:
#R>    $ uvb    : chr  "High" "High" "High" "High" ...
#R>    $ density: Factor w/ 3 levels "1","2","4": 1 1 1 1 1 1 1 1 1 1 ...
#R>    $ mass   : num  0.91 0.65 1.58 0.98 1.07 0.84 0.88 1.47 0.94 0.72 ...
```

An explanatory (or factor) variable must be a character or factor type in R before it can be used in `lm()` for a Two-Way ANOVA.

When performing a Two-Way ANOVA with `lm()` the first argument must be a formula of the form `response~factorA+factorB+factorA:factorB`[1] where `response` is the response variable, `factorA` and `factorB` are the two factor variables, and `factorA:factorB` tells `lm()` to include the interaction between

[1] This exact model may also be entered with the shorthand `response~factorA*factorB`.

the two factor variables. Thus, the Two-Way ANOVA model for the tadpole experiment is fit with

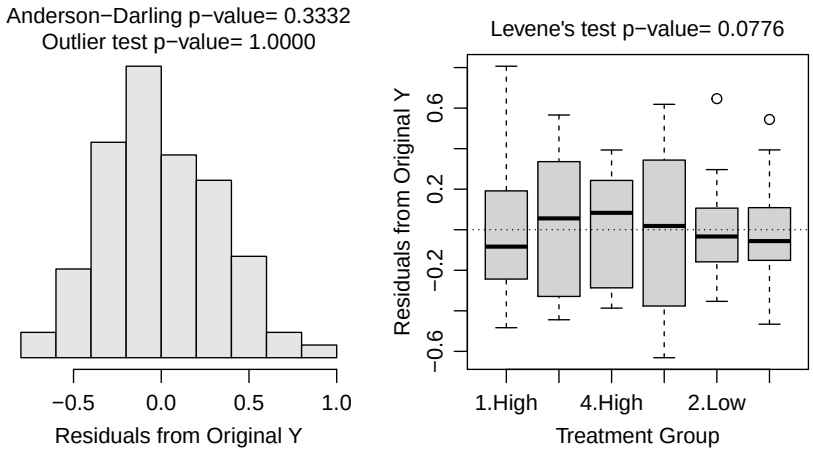
```
lm1 <- lm(mass~density+uvb+density:uvb,data=tad)
```

As usual, the results are saved to an object that will be used to check assumptions, create an ANOVA table, and make multiple comparisons.

12.2 Assumptions

The assumptions for a Two-Way ANOVA are exactly the same as for a One-Way ANOVA as shown in Module 7. Thus, as described in that module, `assumptionCheck()` is used with the saved `lm()` object to compute tests and create graphics to assess the assumptions.

```
assumptionCheck(lm1)
```



From these results it is seen that the variances across all treatments are equal (Levenes p=0.0776), the residuals are normally distributed (p=0.3332), and there are no significant outliers (p>1). There was not much information given previously about this experiment, but as along as the tanks of tadpoles were kept separate such that no tank could impact any other tank then independence is likely adequately met. The analysis can continue with the untransformed data because all assumptions are adequately met.

If the equal variances and normality assumptions (and possibly the no outliers assumption) are not met then try to transform the response variable to a scale

where those assumptions are met. As before, start with the log transformation.

12.3 Main and Interaction Effects (ANOVA Table)

The ANOVA table is extracted from the `lm()` object with `anova()`.

```
anova(lm1)
```

```
#R>   Analysis of Variance Table
#R>
#R>   Response: mass
#R>               Df Sum Sq Mean Sq F value    Pr(>F)
#R>   density      2 2.4337 1.21686 10.3814 9.355e-05
#R>   uvb          1 0.3004 0.30044  2.5632    0.1131
#R>   density:uvb  2 0.2989 0.14947  1.2752    0.2847
#R>   Residuals   84 9.8461 0.11722
```

These results indicate that the interaction effect is insignificant (p=0.2847). Because the interaction term is insignificant, the main effects can be interpreted. The density main effect is strongly significant (p=0.0001), but the UV-B main effect is not significant (p=0.1131). Thus, it appears that the mean body mass of tadpoles differs among some of the density treatments but not between the two UV-B light intensities.

Do not address main effects if there is a significant interaction effect.

12.4 Multiple Comparisons

12.4.1 Main Effects

When an interaction is not present, as is the case here, then multiple comparisons can be conducted for factors related to any main effect that exists. However, multiple comparisons on the main effects are compromised by the interaction term in the model. This is seen by the warning from `emmeans()` below, where I tried to perform multiple comparisons for just the density main effect using `lm1` that contains the interaction term.

```
mc1 <- emmeans(lm1,specs=pairwise~density)
```

Thus, if the interaction term is not significant then you should fit a new model without the interaction term and then use that model when performing multiple comparisons. A model without an interaction term simply uses a formula of the form `response~factorA+factorB` in `lm()`. The ANOVA table is shown below to confirm that the interaction term is not in this new model.

```
lm1_noint <- lm(mass~density+uvb,data=tad)
anova(lm1_noint)
```

```
#R>   Analysis of Variance Table
#R>
#R>   Response: mass
#R>             Df  Sum Sq Mean Sq F value    Pr(>F)
#R>   density    2  2.4337 1.21686 10.3154 9.649e-05
#R>   uvb        1  0.3004 0.30044  2.5469    0.1142
#R>   Residuals 86 10.1450 0.11797
```

To assess main effects with multiple comparisons, first fit (and then use) a model without the insignificant interaction term.

Multiple comparisons for a main effect factor variable are performed as described for a One-Way ANOVA in Section 6.3, but using the new model without the interaction term.

```
mc1_noint <- emmeans(lm1_noint,specs=pairwise~density)
( mc1sum_noint <- summary(mc1_noint,infer=TRUE) )
```

```
#R>   $emmeans
#R>    density emmean     SE df lower.CL upper.CL t.ratio p.value
#R>    1         1.157 0.0627 86    1.033    1.282 18.456  <.0001
#R>    2         0.829 0.0627 86    0.704    0.953 13.215  <.0001
#R>    4         0.791 0.0627 86    0.667    0.916 12.620  <.0001
#R>
#R>   Results are averaged over the levels of: uvb
#R>   Confidence level used: 0.95
#R>
#R>   $contrasts
#R>    contrast estimate     SE df lower.CL upper.CL t.ratio p.value
#R>    1 - 2      0.3287 0.0887 86    0.117    0.540 3.706   0.0011
#R>    1 - 4      0.3660 0.0887 86    0.154    0.578 4.127   0.0002
#R>    2 - 4      0.0373 0.0887 86   -0.174    0.249 0.421   0.9070
#R>
```

```
#R>   Results are averaged over the levels of: uvb
#R>   Confidence level used: 0.95
#R>   Conf-level adjustment: tukey method for comparing a family of 3 estimates
#R>   P value adjustment: tukey method for comparing a family of 3 estimates
```

Again, results for the individual means are in the `$emmeans` portion of the output and the results for differences in paired means are in the `$contrasts` portion. From these results, it is seen that the mean body mass of tadpoles in the 1 tadpole treatment is between 0.117 and 0.540 g greater than the mean for the 2 tadpole treatment (p=0.0011) and between 0.154 and 0.578 g greater than the mean for the 4 tadpole treatment (p=0.0002). The mean body mass of tadpoles did not significantly differ between the 2 and 4 tadpole treatments (p=0.9070).

12.4.2 Interaction Effects

If an interaction effect had been present in the original Two-Way ANOVA model, multiple comparisons must be carried out to determine which pairs of **treatment** means differ. This is easily accomplished with `emmeans()` by using the interaction variable in the `pairwise~` formula. For example, if `lm1` had had a significant interaction term, then multiple comparisons for all pairs of treatments would be computed as follows.

```
mc1 <- emmeans(lm1,specs=pairwise~density:uvb)
( mc1sum <- summary(mc1,infer=TRUE) )
```

```
#R>   $emmeans
#R>    density uvb  emmean     SE df lower.CL upper.CL t.ratio p.value
#R>    1       High  1.023 0.0884 84    0.848    1.199 11.576  <.0001
#R>    2       High  0.784 0.0884 84    0.608    0.960  8.869  <.0001
#R>    4       High  0.797 0.0884 84    0.621    0.972  9.012  <.0001
#R>    1       Low   1.291 0.0884 84    1.116    1.467 14.608  <.0001
#R>    2       Low   0.873 0.0884 84    0.698    1.049  9.879  <.0001
#R>    4       Low   0.786 0.0884 84    0.610    0.962  8.892  <.0001
#R>
#R>   Confidence level used: 0.95
#R>
#R>   $contrasts
#R>    contrast        estimate    SE df lower.CL upper.CL t.ratio p.value
#R>    1 High - 2 High   0.2393 0.125 84  -0.1253   0.6039  1.914  0.4007
#R>    1 High - 4 High   0.2267 0.125 84  -0.1379   0.5913  1.813  0.4631
#R>    1 High - 1 Low   -0.2680 0.125 84  -0.6326   0.0966 -2.144  0.2754
#R>    1 High - 2 Low    0.1500 0.125 84  -0.2146   0.5146  1.200  0.8357
#R>    1 High - 4 Low    0.2373 0.125 84  -0.1273   0.6019  1.898  0.4103
#R>    2 High - 4 High  -0.0127 0.125 84  -0.3773   0.3519 -0.101  1.0000
#R>    2 High - 1 Low   -0.5073 0.125 84  -0.8719  -0.1427 -4.058  0.0015
```

```
#R>    2 High - 2 Low   -0.0893 0.125 84  -0.4539   0.2753 -0.715  0.9797
#R>    2 High - 4 Low   -0.0020 0.125 84  -0.3666   0.3626 -0.016  1.0000
#R>    4 High - 1 Low   -0.4947 0.125 84  -0.8593  -0.1301 -3.957  0.0021
#R>    4 High - 2 Low   -0.0767 0.125 84  -0.4413   0.2879 -0.613  0.9898
#R>    4 High - 4 Low    0.0107 0.125 84  -0.3539   0.3753  0.085  1.0000
#R>    1 Low - 2 Low     0.4180 0.125 84   0.0534   0.7826  3.344  0.0152
#R>    1 Low - 4 Low     0.5053 0.125 84   0.1407   0.8699  4.042  0.0016
#R>    2 Low - 4 Low     0.0873 0.125 84  -0.2773   0.4519  0.699  0.9816
#R>
#R>   Confidence level used: 0.95
#R>   Conf-level adjustment: tukey method for comparing a family of 6 estimates
#R>   P value adjustment: tukey method for comparing a family of 6 estimates
```

As before, results for the individual means are in the `$emmeans` portion of the output and the results for differences in paired means are in the `$contrasts` portion. While this is not an appropriate for this hypothetical set of results, one can see in the results above that the mean body mass for tadpoles in the 4 tadpole and high UVB treatment is between 0.130 and 0.859 less than that for tadpoles in the 1 tadpole and low UVB treatment (p=0.0021).

If a significant interaction term is present, then use multiple comparisons (with the model that has the interaction term) to determine which **treatment** means differ.

12.5 Graphing Results

12.5.1 Interaction Plot

The results of a Two-Way ANOVA are summarizaed in an **interaction plot** whether a significant interaction was present in the results or not. An interaction plot, which was introduced in Module 10, shows each treatment mean with the levels of one factor on the x-axis and the levels of other factor shown with different colors or symbols and sometimes connected with a line. The interaction plots made in this and subsequent modules will also include a confidence intervals for each treatment mean.

The method shown below requires having saved the summary of the multiple comparisons procedure applied to the **model that included the interaction term.** This was created above and saved as `mc1sum`. The data to be plotted is in the `$emmeans` portion of this object (shown below for convenience).

```
mc1sum$emmeans
```

```
#R>    density uvb  emmean     SE df lower.CL upper.CL t.ratio p.value
```

```
#R>    1       High  1.023 0.0884 84    0.848    1.199 11.576  <.0001
#R>    2       High  0.784 0.0884 84    0.608    0.960  8.869  <.0001
#R>    4       High  0.797 0.0884 84    0.621    0.972  9.012  <.0001
#R>    1       Low   1.291 0.0884 84    1.116    1.467 14.608  <.0001
#R>    2       Low   0.873 0.0884 84    0.698    1.049  9.879  <.0001
#R>    4       Low   0.786 0.0884 84    0.610    0.962  8.892  <.0001
#R>
#R>   Confidence level used: 0.95
```

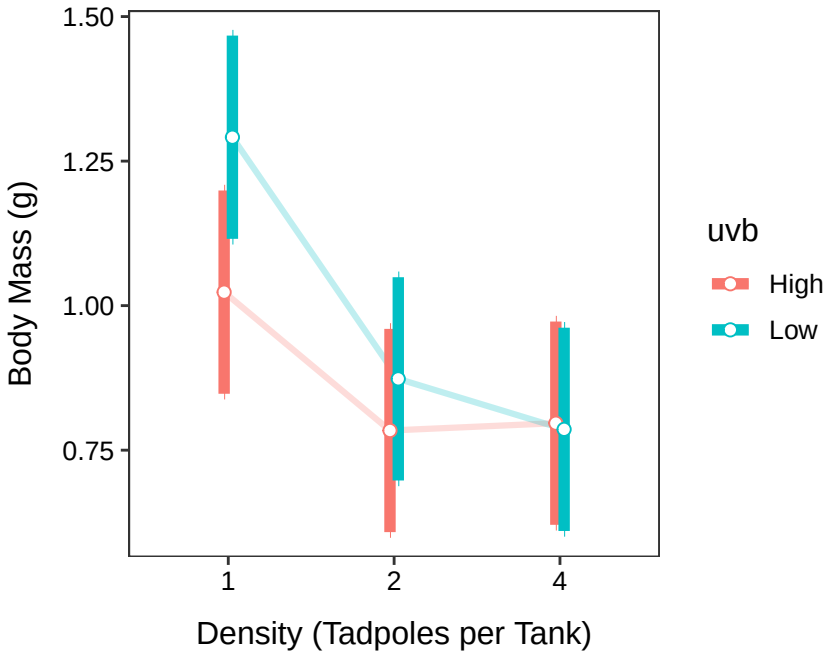
Specifically, we want to plot the results in the `emmean` column with the confidence intervals in `lower.CL` and `upper.CL` against one of the factor variables. One issue that arises is that the confidence intervals for the multiple treatments defined by one level of the variable on the x-axis overlap. Thus, before plotting, a "dodge" amount is defined with `position_dodge()` that will shift the levels slightly left and right to eliminate the overlap. The `width=` argument defines the amount of shift. You may need to "play" with this value to get the exact look that you want.

```
pd <- position_dodge(width=0.1)
```

The summary graphic is constructed with the code below. This code is similar to what was used for the summary graphic of a One-Way ANOVA in Section 6.3. However, this code is somewhat simpler because the individual observations are not plotted to eliminate clutter. Further, note

- the use of `mc1sum$emmeans` as the data,
- the use of `density` (i.e., one of the factor variables) as `x=`,
- the use of `uvb` (i.e., the other factor variale) in `group=` (so it will be "dodged") and `color=` (so it will be denoted with different colors),
- that `geom_line()` is placed first so that the points and confidence intervals will be on top of the connecting lines,
- that `alpha=` is used in `geom_line()` so that the lines are subtle, and
- the use of `pd` from `position_dodge()` above in `geom_pointrange()`.

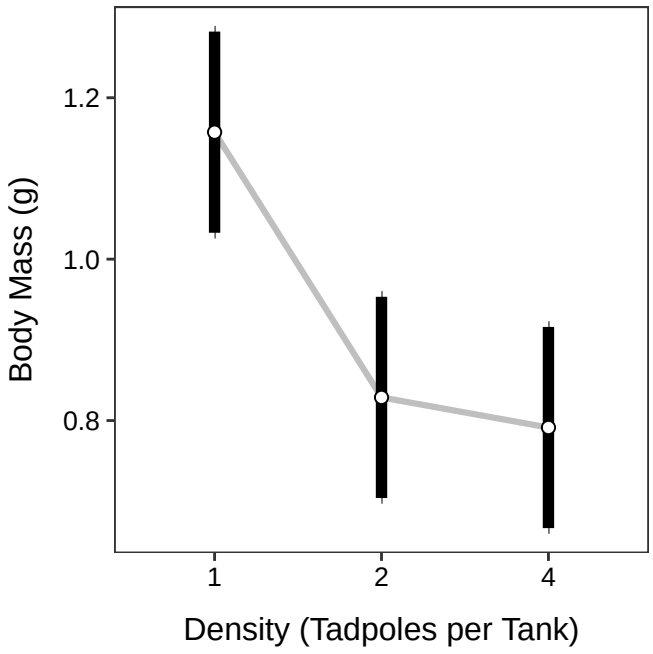
```
ggplot(data=mc1sum$emmeans,mapping=aes(x=density,group=uvb,color=uvb,
                                       y=emmean,ymin=lower.CL,ymax=upper.CL)) +
  geom_line(position=pd,size=1.1,alpha=0.25) +
  geom_errorbar(position=pd,size=2,width=0) +
  geom_point(position=pd,size=2,pch=21,fill="white") +
  labs(y="Body Mass (g)",x="Density (Tadpoles per Tank)") +
  theme_NCStats()
```



12.5.2 Main Effects Plot

Some researchers prefer to plot just the main effects when there is no significant interaction effect in the data. Such a plot is called a **main effects** plot and can be constructed similarly from the multiple comparison results using the model without an interaction term. Note that `group=1` must be used as shown below so that `geom_line()` will work properly.

```
ggplot(data=mc1sum_noint$emmeans,
       mapping=aes(x=density,group=1,y=emmean,ymin=lower.CL,ymax=upper.CL)) +
  geom_line(size=1.1,alpha=0.25) +
  geom_errorbar(size=2,width=0) +
  geom_point(size=2,pch=21,fill="white") +
  labs(y="Body Mass (g)",x="Density (Tadpoles per Tank)") +
  theme_NCStats()
```


1.2
1.0
0.8
Body Mass (g)
1
2
4
Density (Tadpoles per Tank)

Chapter 13

Two-Way Summary

Specific parts of a full Two-Way ANOVA analysis were described in Module 12. In this module, a workflow for a full analysis is offered and that workflow is demonstrated with several examples.

13.1 Suggested Workflow

The following is a process for fitting a Two-Way ANOVA model. Consider this process as you learn to fit Two-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 special attention to the distributional shape, center, dispersion, and outliers within each treatment/group.
2. Show the sample size per group and comment on whether the study was balanced (i.e., same sample size per group) or not.
3. Address the independence assumption.
    - If this assumption is not met then other analysis methods must be used.
4. Fit the untransformed ultimate full model (i.e., both main effects and the interaction effect) model with `lm()`.
5. Check the other three assumptions for the untransformed model with `assumptionCheck()`.
    - Check equality of variances with a Levene's test and residual plot.
    - Check normality of residuals with a Anderson-Darling test and histogram of residuals.
    - Check for outliers with an outlier test, residual plot, and histogram of residuals.
6. If an assumption or assumptions are violated, then attempt to find a transformation where the assumptions are met.

- Use the trial-and-error method with `assumptionCheck()`, theory, or experience to identify a possible transformation. Always try the log transformation first.
- If only an outlier exists (i.e., there are equal variances and normal residuals) and no transformation corrects the outlier then consider removing the outlier from the data set.
- Fit the ultimate full model with the transformed response or reduced data set.

7. Construct an ANOVA table for the full model with `anova()`.
   - If a significant interaction exists then do NOT interpret the main effects!!
   - If a significant interaction does NOT exist then interpret the main effects.
     - Fit a new model without the insignificant interaction term.
8. If an effect exists, then use a multiple comparison technique with `emmeans()` and `summary()` to identify specific differences. Describe specific differences using confidence intervals.
   - If a significant interaction exists then perform multiple comparisons on the interaction term using the model that contained an interaction term.
   - If a significant interaction does not exist then perform multiple comparisons for each factor for which a main effect exists using the model without an interaction term.
9. Create a summary graphic of treatment means (i.e., an interaction plot) **on the original scale** with 95% confidence intervals using `ggplot()` and results from `emmeans()` using the model with an interaction term.
10. Write an overall conclusion of your findings.

13.2 Expected Prices (*No Transformation*)

Managers of a retail store felt that the price that consumers would expect to pay for a product would be influenced by how much the product was promoted and the advertised amount of discount. To examine this, they gathered 160 volunteers (from different households) who would receive information about the store's products for a 10-week period. Each volunteer was randomly chosen to receive promotions about one particular product 1, 3, 5, or 7 times during that period and with an advertised discount of 10, 20, 30, or 40%. At the end of the 10-week period the researchers asked each participant to report the price they would expect to pay for the product.[1]

The statistical hypotheses to be examined were

[1] This example is modified from Alwyn *et al.* (2020) who based it on Kalwani and Kim (1992).

Table 13.1: Number of participants in each combination of number of promotions and amount of discount.

	Discounts			
	10	20	30	40
Promotions				
1	10	10	10	10
3	10	10	10	10
5	10	10	10	10
7	10	10	10	10

$$H_0 : \text{no promotions effect} : \mu_1 = \mu_3 = \mu_5 = \mu_7$$
$$H_A : \text{promotions effect} : \text{At least one pair of level means is different}$$

$$H_0 : \text{no discount effect} : \mu_{10} = \mu_{20} = \mu_{30} = \mu_{40}$$
$$H_A : \text{discount effect} : \text{At least one pair of level means is different}$$

$$H_0 : \text{no interaction effect}$$
$$H_A : \text{interaction effect}$$

where μ is the mean expected price and the subscripts identify the levels of each factor as defined above.

This study was "balanced" as the number of participants was the same in each combination of number of promotions and amount of discount (Table 13.1).

Variances among the treatments appear to be constant (Levene's p=0.4162) and the boxplots of residuals appear fairly similar (Figure 13.1-Right); the residuals appear to be approximately normally distributed (Anderson-Darling p=0.6255; Figure 13.1-Left); and there are no significant outliers (outlier test p=0.6285), though some residuals appear somewhat larger in some treatments (Figure 13.1-Right).

The participants were not randomly selected for the study but they were specifically not from the same household and they were randomly allocated to the combination of number of promotions and discount amount. Thus, there is no reason to believe that individuals are connected either within or among treatments. Thus, the independence assumption appears to have also been met. These data will be examined with a Two-Way ANOVA without transformation because the assumptions have been adequately met.

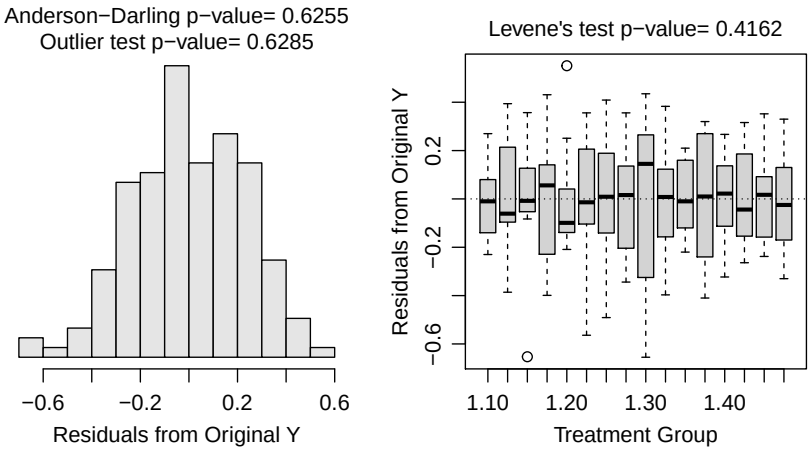


Figure 13.1: Histogram of residuals (Left) and residual plot (Right) for a Two-way ANOVA of expected price for each combination of number of promotions and discount rate.

Table 13.2: Two-way ANOVA results for expected price among different numbers of promotions and amounts of discount.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Promo	3	6.024	2.008	34.385	0.0000
Discount	3	7.824	2.608	44.661	0.0000
Promo:Discount	9	0.231	0.026	0.439	0.9121
Residuals	144	8.409	0.058		

There does not appear to be a significant interaction effect (p=0.9121; Table 13.2). There does however appear to be main effects for both the number of promotions (p<0.00005) and amount of discount (p<0.00005).

Tukey's multiple comparison results suggest that the mean expected price does not differ between when one and three promotions are used (p=0.9476) but does decline from three to five promotions (p<0.00005) and from five to seven promotions (p=0.0079; Table 13.3). For example, the mean expected price drops between 0.13 and 0.41 dollars from three to five promotions and between 0.03 and 0.31 dollars from five to seven promotions.

Tukey's multiple comparison results suggest that the mean expected price differed between all pairs of amounts of discount (p 0.0021; Table 13.4). The mean expected price declined between 0.06 and 0.33 dollars from a discount rate of 10% to 20% (p=0.0019) and between 0.26 and 0.53 dollars from a discount rate of 20% to 30% (p<0.00005), but increased between 0.06 and 0.33

Table 13.3: Tukey's multiple comparisons of differences in mean expected price among all pairs of numbers of promotions.

contrast	estimate	lower.CL	upper.CL	p.value
1 - 3	0.03	-0.11	0.17	0.9476
1 - 5	0.30	0.16	0.43	0.0000
1 - 7	0.47	0.33	0.61	0.0000
3 - 5	0.27	0.13	0.41	0.0000
3 - 7	0.44	0.30	0.58	0.0000
5 - 7	0.17	0.03	0.31	0.0079

Table 13.4: Tukey's multiple comparisons of differences in mean expected price among all pairs of amounts of discount.

contrast	estimate	lower.CL	upper.CL	p.value
10 - 20	0.19	0.06	0.33	0.0019
10 - 30	0.59	0.45	0.73	0.0000
10 - 40	0.40	0.26	0.54	0.0000
20 - 30	0.40	0.26	0.53	0.0000
20 - 40	0.20	0.07	0.34	0.0011
30 - 40	-0.19	-0.33	-0.06	0.0021

dollars from a discount rate of 30% to 40% (p=0.0021).

From these results it appears that if there is no difference in mean expected price for three or less promotions, but after that the mean expected price drops with increasing numbers of promotions, regardless of the amount of discount offered. Regardless of the number of promotions, consumers expect to pay less with increasing amount of discount up to 30%, but the mean expected price increased when the discount was increased to 40% (Figure 13.2).

R Code and Results

```
d <- read.csv("http://derekogle.com/Book207/data/Discount.csv")
d$Promo <- factor(d$Promo)
d$Discount <- factor(d$Discount)
lm1.d <- lm(Eprice~Promo+Discount+Promo:Discount,data=d)
xtabs(~Promo+Discount,data=d)
assumptionCheck(lm1.d)
```

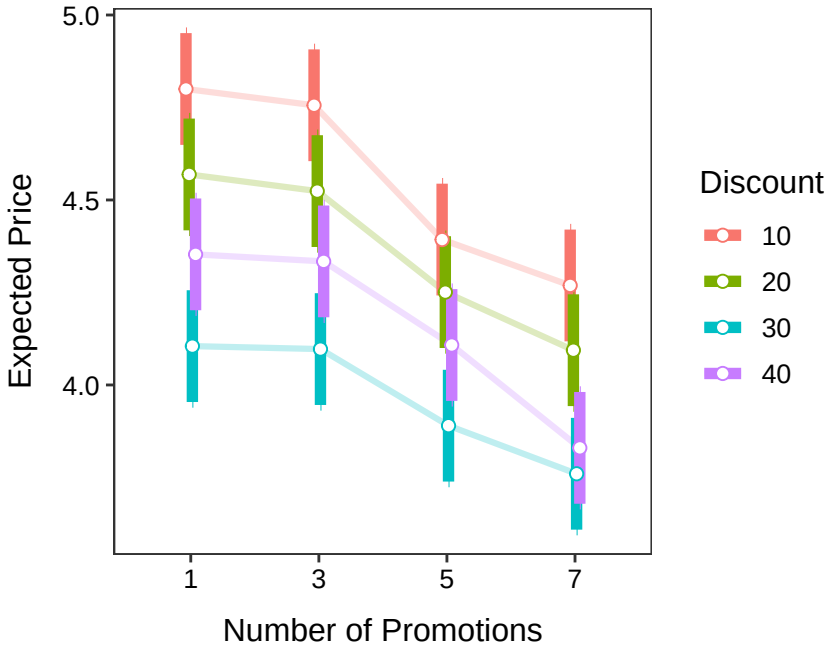


Figure 13.2: Mean expected price for each combination of number of promotions and discount amount.

```
anova(lm1.d)
lm1.d.noint <- lm(Eprice~Promo+Discount,data=d)
mc1.d.promo <- emmeans(lm1.d.noint,specs=pairwise~Promo)
( mc1sum.d.promo <- summary(mc1.d.promo,infer=TRUE) )
mc1.d.discount <- emmeans(lm1.d.noint,specs=pairwise~Discount)
( mc1sum.d.discount <- summary(mc1.d.discount,infer=TRUE) )
```

13.3 Blood Pressure (*No Transformation*)

Sodium (Na) plays an important role in the genesis of high blood pressure, and the kidney is the principal organ that regulates the amount of sodium in the body. The kidney contains the Na-K-ATPase enzyme, which is essential for maintaining proper sodium levels. If the enzyme does not function properly, then high blood pressure may result. The activity of this enzyme has been studied in whole kidneys, even though the kidney is known to contain many functionally distinct sites. To see whether any particular site of Na-K-ATPase activity was abnormal with hypertension, Garg *et al.* (1985) studied Na-K-ATPase activity at different sites along the nephrons of normal rats and specially-bred rats which spontaneously develop hypertension. The sites in the kidney examined were the distal collecting tubule (DCT), cortical collecting duct (CCD), and outer medullary collecting duct (OMCD).

The researchers hypothesized that the level of Na-K-ATPase would be depressed at all sites in the rats with hypertension. This translates to expecting a rat strain effect but not an interaction effect. The authors were also interested in determining if there was a significant difference in the level of Na-K-ATPase activity at the different sites. Thus, the statistical hypotheses to be examined were

$$H_0 : \text{no rat strain effect} : \mu_{Normal} = \mu_{Hyper}$$
$$H_A : \text{rat strain effect} : \mu_{Normal} \neq \mu_{Hyper}$$

$$H_0 : \text{no site effect} : \mu_{DCT} = \mu_{CCD} = \mu_{OMCD}$$
$$H_A : \text{site effect} : \text{At least one pair of level means is different}$$

$$H_0 : \text{no interaction effect}$$
$$H_A : \text{interaction effect}$$

Table 13.5: Number of rats in each combination of rat strain and kidney site.

	Kidney Site		
	CCD	DCT	OMCD
Strain			
Hyper	4	4	4
Normal	4	4	4

where μ is the mean Na-K-ATPase activity and the subscripts identify the levels of each factor as defined above.

Twelve rats were randomly selected for each strain of rat (i.e., normal rats and rats with hypertension) from all rats available in the author's laboratory. The site of the kidney (DCT, CCD, or OMCD) where the Na-K-ATPase activity ($pmol \cdot (min \cdot mm)^{-1}$) was recorded on a rat was also randomly selected so that measurements at each location were recorded from four rats.[2]

Variances among treatments appear to be constant (Levene's p=0.4033) though the boxplots of residuals appear somewhat divergent (Figure 13.3-Right), likely due to the small number of rats per treatment; the residuals appear to be approximately normally distributed (Anderson-Darling p=0.5203; Figure 13.3-Left); and there are no significant outliers (outlier test p=0.1527).

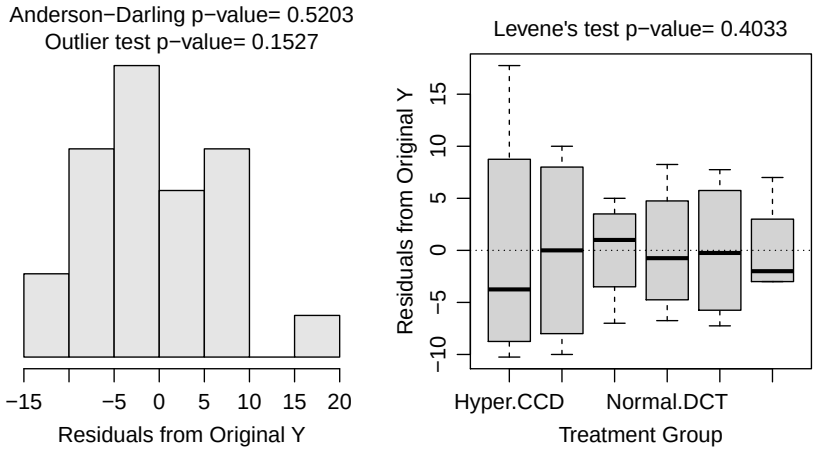


Figure 13.3: Histogram of residuals (Left) and residual plot (Right) for a Two-way ANOVA of Na-K-ATPase activity for each combination of rat type and measurement location.

[2]A better design would have measured the Na-K-ATPase activity at all three sites from the same rates. However, this would have violated the independence assumption and required other methods (repeated-measures or mixed-models) to analyze the data.

Table 13.6: Two-way ANOVA results for Na-K-ATPase activity levels among two strains of rats and three sites in the kidney.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
strain	1	459.4	459.4	7.139	0.0156
site	2	8287.0	4143.5	64.393	0.0000
strain:site	2	496.0	248.0	3.854	0.0404
Residuals	18	1158.2	64.3		

The rats are probably independent among strains as there is no evidence that "normal" and "hypertensive" rats are in any way related. Independence of rats within strains is suspect given that the rats came from the same "pool" of normal and hypertensive rats bred in the lab. This may be acceptable for this experiment, but drawing conclusions to a larger population of rats may be suspect. Finally, rats are likely independent among kidney site groups as a different site was examined for each individual rat. All-in-all, independence is likely met enough to continue with this analysis. Thus, these data will be examined with a two-way ANOVA without transformation because the assumptions have been adequately met.

There appears to be a (weakly) significant interaction effect (p=0.0404; Table 13.6); thus, the main effects cannot be interpreted directly from these results.

Tukey's method was performed on the interaction effect to more specifically describe the differences among group means (Table 13.7). The mean level of Na-K-ATPase activity was significantly lower for the hypertensive rats then for the normal rats at the DCT site (p=0.0189), but not at the other two sites (p 0.8359). For example, the mean level of Na-K-ATPase activity was between 2.7 and 38.8 units lower in hypertensive than normal rats at the DCT site. In addition, the mean level of Na-K-ATPase activity was significantly greater at the DCT site than at the CCD and OMCD sites within both strains (p 0.0029), and Na-K-ATPase activity did not differ between the CCD and OMCD sites for either strain (p 0.1367). For example, the mean level of Na-K-ATPase activity was between 36.7 and 72.8 units higher at the DCT than at the OMCD site for normal rats.

It can be difficult to interpret multiple comparisons when an interaction exists, especially if there are many differences among treatment means. A strategy to handle this is to describe how one factor differs among levels of the other factor (e.g., how strain differed among sites) and then vice-versa (e.g., how sites differed among strains).

There is some support for the researcher's hypothesis that Na-K-ATPase activity levels would be lower in rats with hypertension (Figure 13.4).

Table 13.7: Tukey's multiple comparisons of differences in means among all pairs of combinations of rat strain and kidney measurement site.

contrast	estimate	lower.CL	upper.CL	p.value
CCD Hyper - DCT Hyper	-25.75	-43.78	-7.72	0.0029
CCD Hyper - OMCD Hyper	7.00	-11.03	25.03	0.8148
CCD Hyper - CCD Normal	-6.75	-24.78	11.28	0.8359
CCD Hyper - DCT Normal	-46.50	-64.53	-28.47	0.0000
CCD Hyper - OMCD Normal	8.25	-9.78	26.28	0.6953
DCT Hyper - OMCD Hyper	32.75	14.72	50.78	0.0002
DCT Hyper - CCD Normal	19.00	0.97	37.03	0.0355
DCT Hyper - DCT Normal	-20.75	-38.78	-2.72	0.0189
DCT Hyper - OMCD Normal	34.00	15.97	52.03	0.0001
OMCD Hyper - CCD Normal	-13.75	-31.78	4.28	0.1997
OMCD Hyper - DCT Normal	-53.50	-71.53	-35.47	0.0000
OMCD Hyper - OMCD Normal	1.25	-16.78	19.28	0.9999
CCD Normal - DCT Normal	-39.75	-57.78	-21.72	0.0000
CCD Normal - OMCD Normal	15.00	-3.03	33.03	0.1367
DCT Normal - OMCD Normal	54.75	36.72	72.78	0.0000

However, this support was found only for the distal collecting tubule (DCT) site. At all other sites, no difference between hypertensive and normal rats was observed. Furthermore, Na-K-ATPase activity was higher at the DCT site then either of the other two sites for both normal and hypertensive rats.

R Code and Results

```
nak <- read.csv("http://derekogle.com/Book207/data/NAKATPase.csv")
lm1.nak <- lm(activity~strain+site+strain:site,data=nak)
xtabs(~strain+site,data=nak)
assumptionCheck(lm1.nak)
```

```
anova(lm1.nak)
mc1.nak <- emmeans(lm1.nak,specs=pairwise~site:strain)
( mc1sum.nak <- summary(mc1.nak,infer=TRUE) )
pd <- position_dodge(width=0.1)
ggplot(data=mc1sum.nak$emmeans,
       mapping=aes(x=site,group=strain,color=strain,
                   y=emmean,ymin=lower.CL,ymax=upper.CL)) +
  geom_line(position=pd,size=1.1,alpha=0.25) +
  geom_errorbar(position=pd,size=2,width=0) +
```

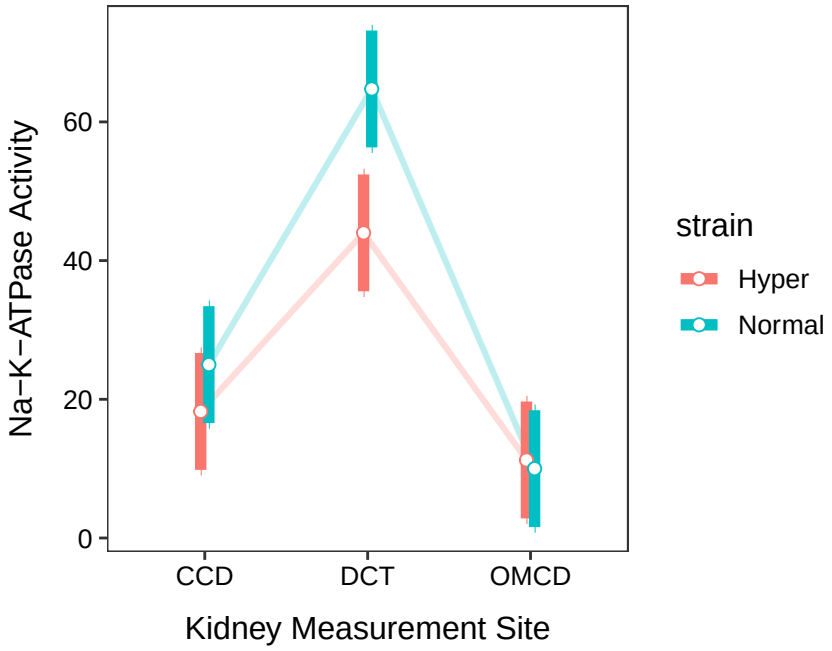


Figure 13.4: Mean Na-K-ATPase activity for each combination of rat strain and kidnety measurement site.

```
  geom_point(position=pd,size=2,pch=21,fill="white") +
  labs(y="Na-K-ATPase Activity",x="Kidney Measurement Site") +
  theme_NCStats()
```

13.4 Crayfish Foraging (*Transformation*)

Nystrom and Graneli (1996) examined the importance of intraspecific competition for food as a factor regulating survival, growth, and fecundity in Noble Crayfish (*Astacus astacus*). In one aspect of their research they examined factors that increased the risk of predation. In this part of their study, the authors assumed that the number of active crayfish (i.e., not in shelters) was an indicator of predation risk (i.e., the crayfish are out of their shelters and are thus more vulnerable). The authors hypothesized that the crayfish's willingness to risk predation would be greater if competition for food was greater. Thus, the authors created two levels of "competition" by regulating how much food the crayfish received (with the assumption that competition is greater with lesser food). The two feeding regimes were that crayfish were fed *ad libitum* and that they were fed slightly less than a maintenance ration (called "unfed"). In addition, there is ample evidence that crayfish are more active near dusk and at night as darkness provides some protection from predatory fish. To test for this, the authors included a time of day factor in their study, that had three levels: 1200 (noon), 1700, and 1900.

The authors had a large number of crayfish collected from a local lake that were available to them for this experiment. They randomly placed the crayfish into groups of 50 crayfish that were then "stocked" into plastic tubs that had been filled with sand, small pebbles, and artificial shelters. The tubs were placed haphazardly around an outside area exposed to natural light. Each tub was then assigned a "treatment" that consisted of a combination of feeding levels and the time of day when the number of active crayfish (out of 50) that were active (i.e., not in a shelter) would be recorded.

The statistical hypotheses to be examined were

$$H_0 : \text{no competition effect} : \mu_{Fed} = \mu_{Unfed}$$
$$H_A : \text{competition effect} : \mu_{Fed} \neq \mu_{Unfed}$$

$$H_0 : \text{no time effect} : \mu_{1200} = \mu_{1700} = \mu_{1900}$$
$$H_A : \text{time effect} : \text{At least one pair of level means is different}$$

Table 13.8: Number of groups of crayfish in each combination of feeding regime and time of day when the number of active crayfish were recorded.

	Kidney Site		
	12	17	19
Strain			
Fed	6	12	6
Unfed	6	12	6

$$H_0 : \text{no interaction effect}$$
$$H_A : \text{interaction effect}$$

where μ is the mean number of active crayfish and the subscripts identify the levels of each factor as defined above.

This study is not balanced (Table 13.8) as the authors chose to record more groups of crayfish at the 1700 (i.e., dusk) time.

The authors appeared to take steps to meet the independence assumption. There appears to be independence across all treatments as they used a different set of 50 crayfish for each treatment. In addition, the tubs were placed haphazardly around the area so there should not be any spatial effect such as all unfed treatments being in the same area. Furthermore, they only recorded the number of active crayfish at one time for each tub of crayfish. It would have been much easier to record all three times for each tub of crayfish but this would have violated the independence assumption and required a different analytical method.

Variances among the treatments appear not to be constant (Levene's p=0.0201) and the boxplots of residuals are quite divergent (Figure 13.5-Right); the residuals do not appear to be approximately normally distributed and are quite left-skwed (Anderson-Darling p<0.00005; Figure 13.5-Left); and there are significant outliers (outlier test p=0.0002; Figure 13.5). Clearly the assumptions are not met and a transformation should be considered.

No transformation worked perfectly for these data. However, a log transformation resulted in equal variances (Levene's p=0.6259), with boxplots of residuals that are not wildly divergent (Figure 13.6-Right); the residuals do not appear to be approximately normally distributed (p=0.0004), but are not too strongly skewed (Figure 13.6-Left); and there are still significant outliers (outlier test p=0.0104; Figure 13.6), mostly in the Unfed-1700 group. I did not remove outliers as the sample size is already quite small. The assumptions are

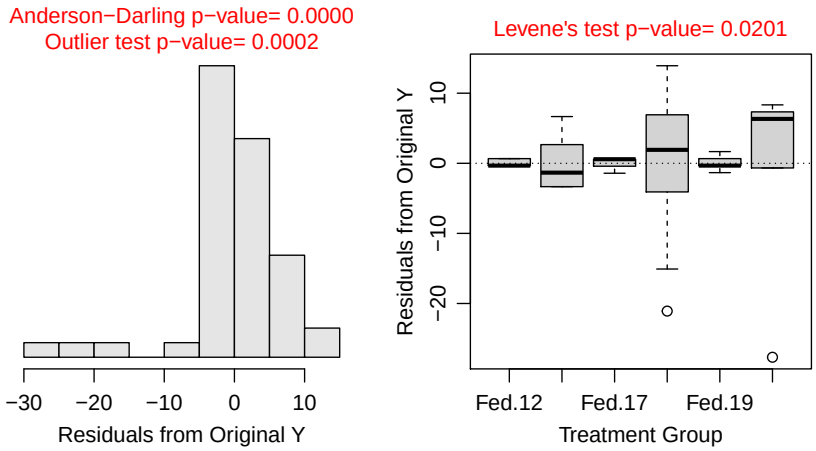


Figure 13.5: Histogram of residuals (Left) and residual plot (Right) for a Two-way ANOVA of the number of active crayfish (out of 50) activity for each combination of feeding level and time of day.

Table 13.9: Two-way ANOVA results for number of active crayfish for two feeding levels and three times of day.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feed	1	42.12	42.12	236.46	0.00000
time	2	7.73	3.86	21.69	0.00000
feed:time	2	3.14	1.57	8.81	0.00064
Residuals	42	7.48	0.18		

more closely met on the log scale, so I will continue to analyze the data with this transformation.

There appears to be a significant interaction effect (p=0.0006; Table 13.9); thus, the main effects cannot be interpreted directly from these results.

Tukey's multiple comparisons (Table 13.10) show that mean number of active crayfish did not differ by time of day when the crayfish were fed *ad libitum* (p 0.5042). In contrast, in the unfed treatments, the mean number of active crayfish increased from 1200 to 1700 (p<0.00005), but not from 1700 to 1900 (p=0.3489). Additionally, the mean number of active fish was greater in the unfed than in the fed treatments for each time of day (p 0.0013). For example, the mean number of active crayfish in the unfed treatment was between 5.63 and 24.12 **times** greater than the mean number of active crayfish in the fed treatment at 1900.

The researcher's primary hypotheses was supported by this study – crayfish in

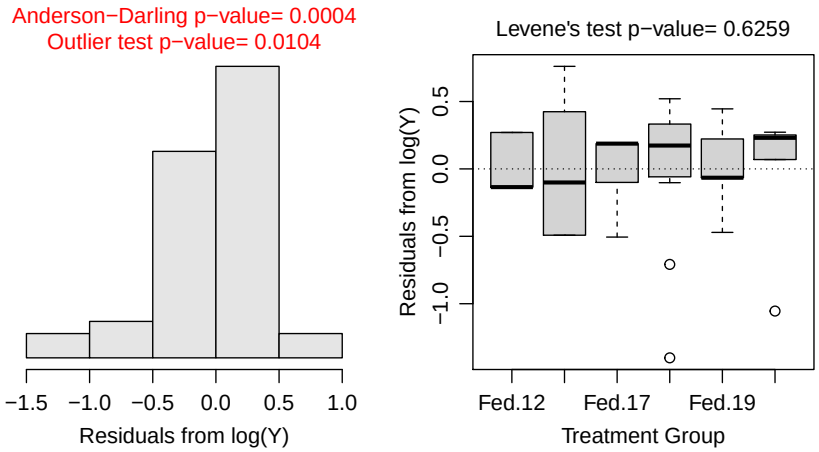


Figure 13.6: Histogram of residuals (Left) and residual plot (Right) for a Two-way ANOVA of the number of active crayfish (out of 50) activity for each combination of feeding level and time of day.

Table 13.10: Tukey's multiple comparisons for **ratios** of means among all pairs of combinations of feeding levels and times of day. Results were back-transformed from the log scale.

contrast	ratio	lower.CL	upper.CL	p.value
Fed 12 / Unfed 12	0.35	0.17	0.72	0.0013
Fed 12 / Fed 17	0.69	0.37	1.30	0.5042
Fed 12 / Unfed 17	0.09	0.05	0.18	0.0000
Fed 12 / Fed 19	0.71	0.35	1.48	0.7396
Fed 12 / Unfed 19	0.06	0.03	0.13	0.0000
Unfed 12 / Fed 17	1.97	1.05	3.70	0.0279
Unfed 12 / Unfed 17	0.27	0.14	0.50	0.0000
Unfed 12 / Fed 19	2.04	0.99	4.23	0.0570
Unfed 12 / Unfed 19	0.18	0.08	0.36	0.0000
Fed 17 / Unfed 17	0.14	0.08	0.23	0.0000
Fed 17 / Fed 19	1.04	0.55	1.94	1.0000
Fed 17 / Unfed 19	0.09	0.05	0.17	0.0000
Unfed 17 / Fed 19	7.61	4.05	14.29	0.0000
Unfed 17 / Unfed 19	0.65	0.35	1.23	0.3489
Fed 19 / Unfed 19	0.09	0.04	0.18	0.0000

a high competition situation (i.e., unfed) which more willing to take on predation risk by being active outside their shelters (Figure 13.7). Their thought that crayfish would be more active near dusk or at night was partially supported as this was only evident when crayfish were in the high competition scenario.

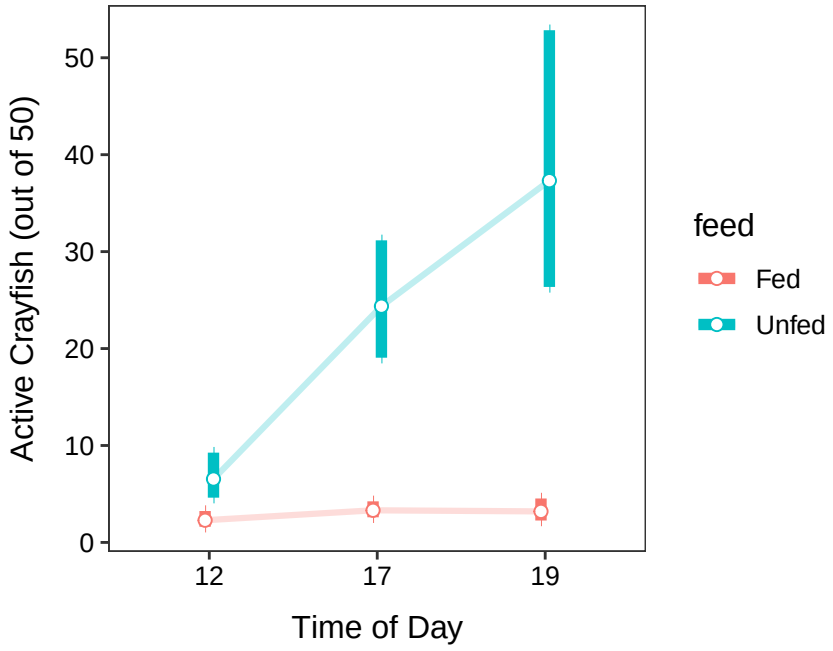


Figure 13.7: Mean number of active crayfish for each combination of feeding rate and time of day. Means and confidence intervals are back-transformed from the log scale.

R Code and Results

```
cray <- read.csv("http://derekogle.com/Book207/data/CrayfishCompetition.csv")
cray$feed <- factor(cray$feed)
cray$time <- factor(cray$time)
lm1.cc <- lm(active~feed+time+feed:time,data=cray)
xtabs(~feed+time,data=cray)
cray$logact <- log(cray$active)
lm1.cct <- lm(logact~feed+time+feed:time,data=cray)
assumptionCheck(lm1.cc,lambday=0)
```

```
anova(lm1.cct)
mc1.cct <- emmeans(lm1.cct,specs=pairwise~feed:time,tran="log",type="response")
( mc1sum.cct <- summary(mc1.cct,infer=TRUE) )
pd <- position_dodge(width=0.1)
ggplot(data=mc1sum.cct$emmeans,
       mapping=aes(x=time,group=feed,color=feed,
                   y=response,ymin=lower.CL,ymax=upper.CL)) +
  geom_line(position=pd,size=1.1,alpha=0.25) +
  geom_errorbar(position=pd,size=2,width=0) +
  geom_point(position=pd,size=2,pch=21,fill="white") +
  labs(y="Active Crayfish (out of 50)",x="Time of Day") +
  theme_NCStats()
```