Readings for MTH207

Derek H. Ogle

2020-12-20

Contents

Preface

This BOOK-IN-PROGRESS was last updated on: 20 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

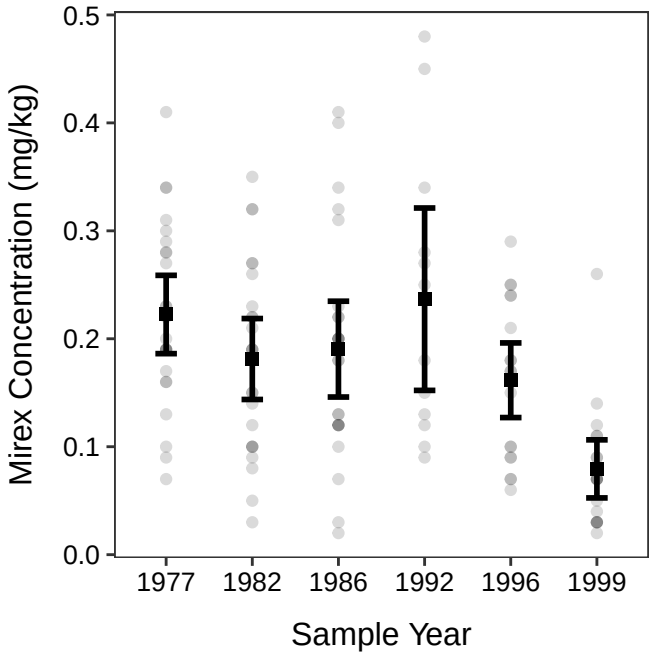


Figure 1.1: Mean mirex concentration by sample year. This is an example of a One-Way ANOVA.

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.

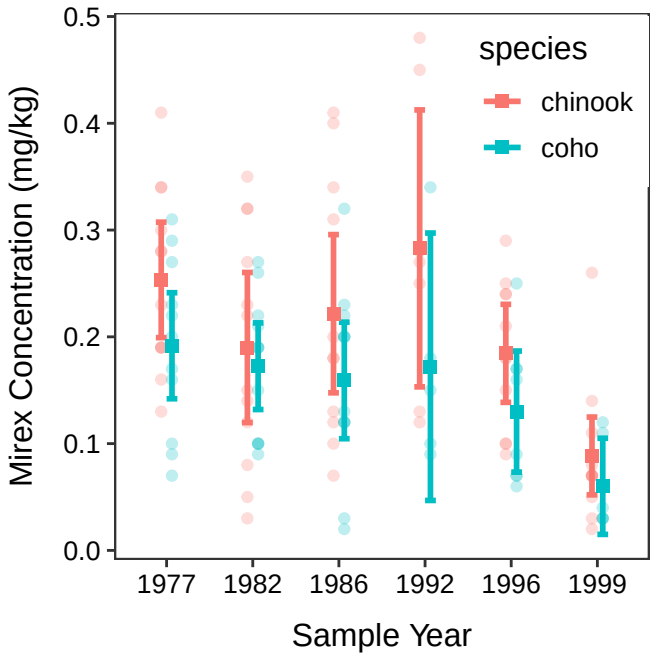


Figure 1.2: Mean mirex concentration by sample year and salmon species. This is an example of a Two-Way ANOVA.

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".

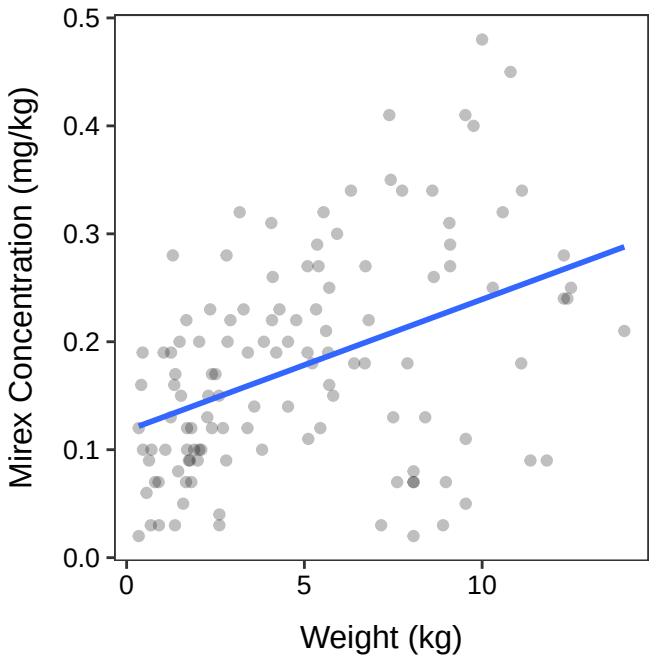


Figure 1.3: Mirex concentration by fish weight. This is an example of a Simple Linear Regression.

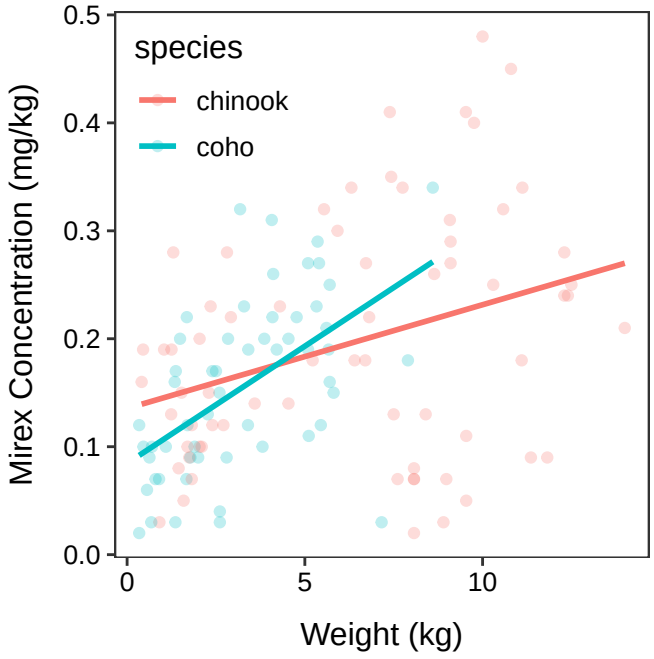


Figure 1.4: Mirex concentration by fish weight seprated by salmon species. This is an example of an Indicator Variable Regression.

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

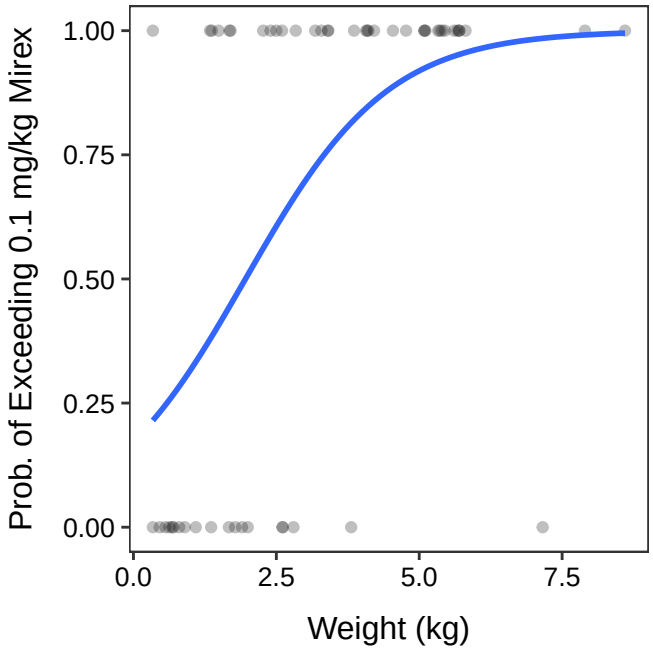


Figure 1.5: The probability that the mirex concentration exceed the 0.1 mg/kg threshold by fish weight. This is an example of a Logistic Regression.

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.

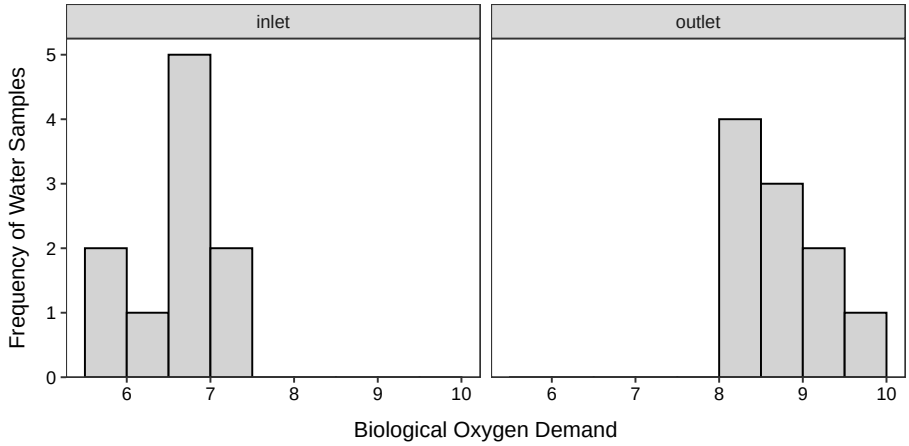
Specific details for performing a 2-sample t-test in R were provided in your introductory statistics course, but will be cursorily reviewed below.

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.

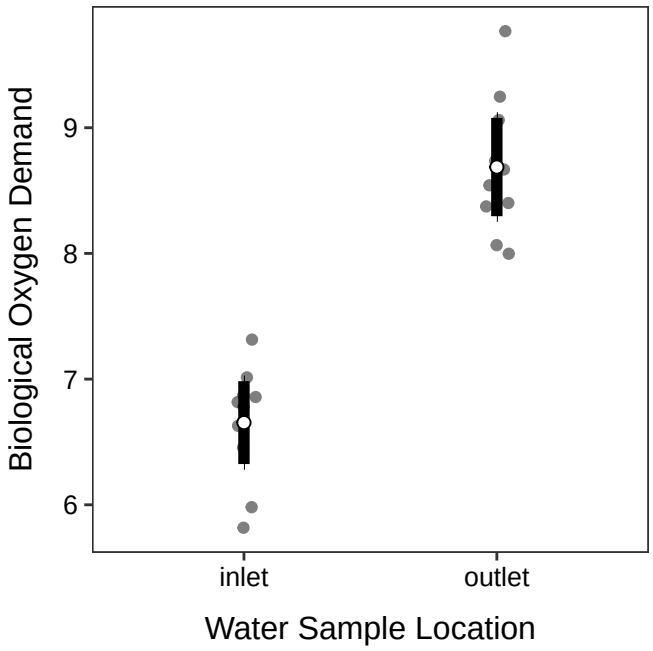
```
t.test(BOD~src,data=aqua,var.equal=TRUE)
```

```
#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

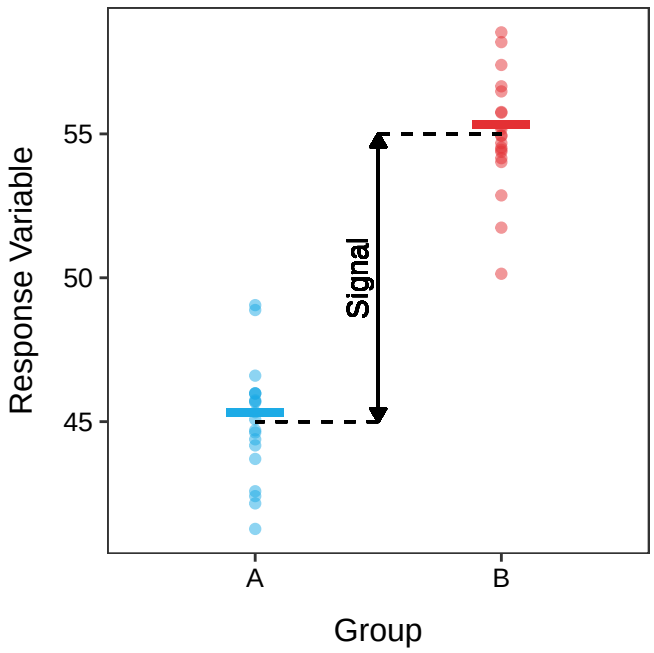


Figure 2.1: Response variable by group for each indiviual (points) with group means shown as horizontal segments. The difference in sample means is highlighted as the "signal" in these data.

"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

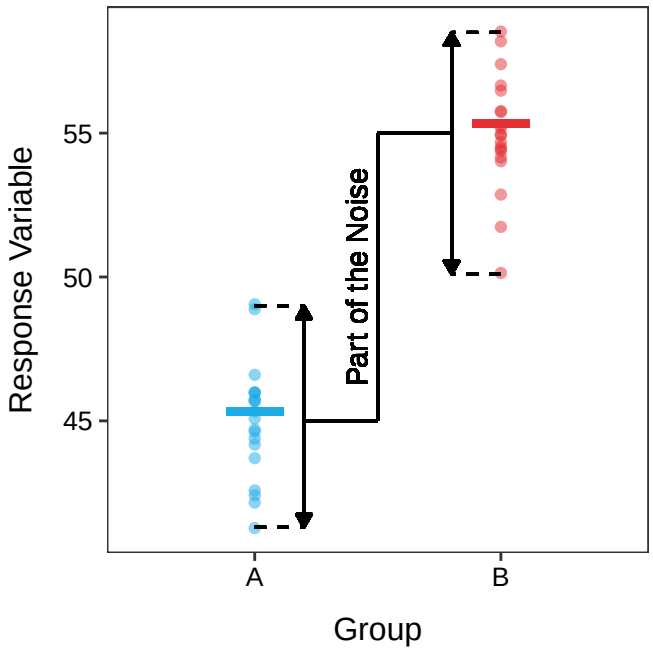


Figure 2.2: Response variable by group for each indiviual (points) with group means shown as horizontal segments. The variability of individuals around the group means is highlighted as a part of the "noise" in these data.

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).

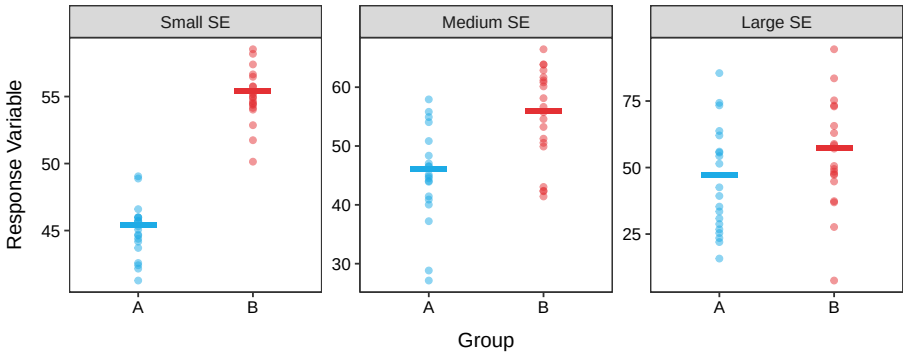


Figure 2.3: Response variable by group for each indiviual (points) with group means shown as horizontal segments for three different standard errors (SE; i.e., "noise"). Note that the group means are the same in all three panels.

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

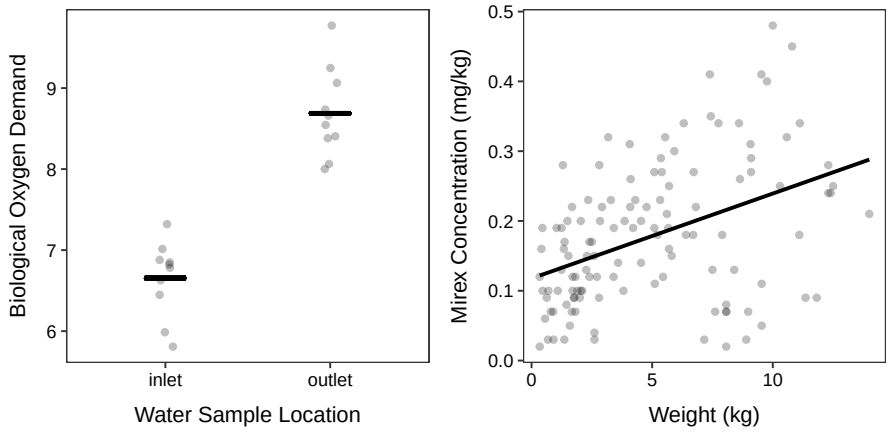


Figure 3.1: Two examples of models seen in your introductory statistics course – two means (Left) and regression line (Right).

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.

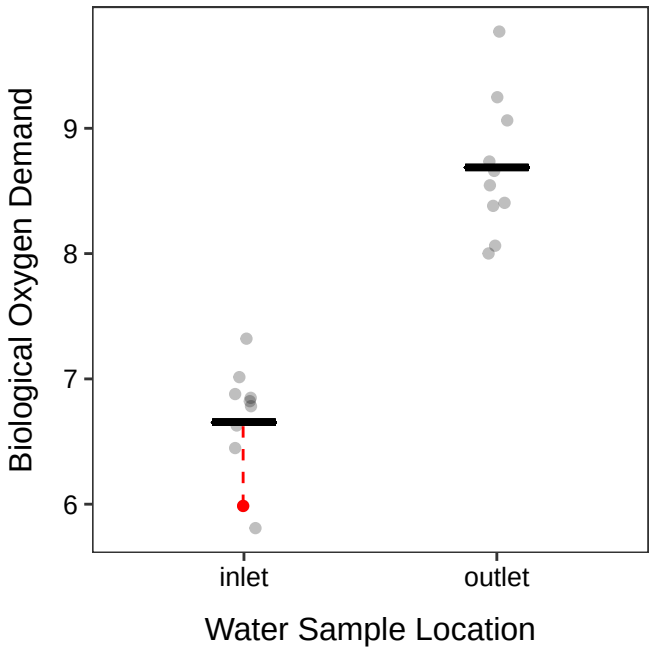


Figure 3.2: Biological oxygen demand versus sample location (points) with group means shown by horizontal segments. The residual from a model that uses a separate mean for both groups is shown.

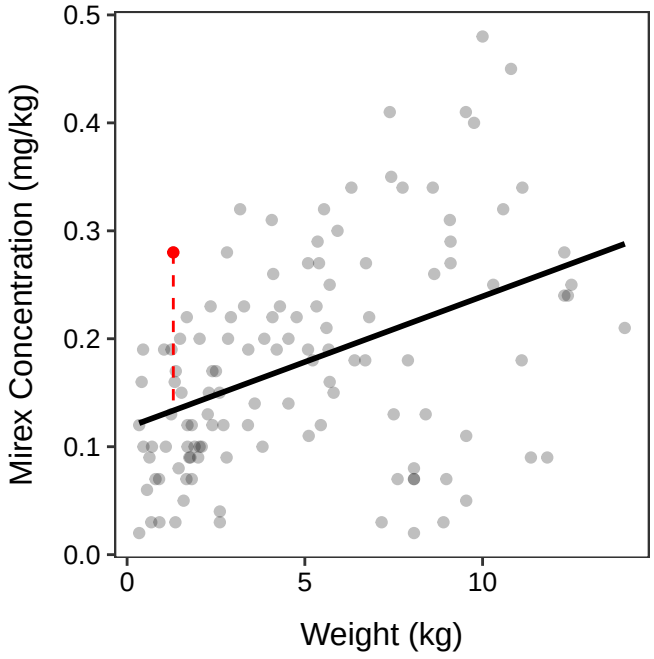


Figure 3.3: Mirex concentration versus fish weight with a simple linear regression line show. The residual from the regression line model is shown.

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.

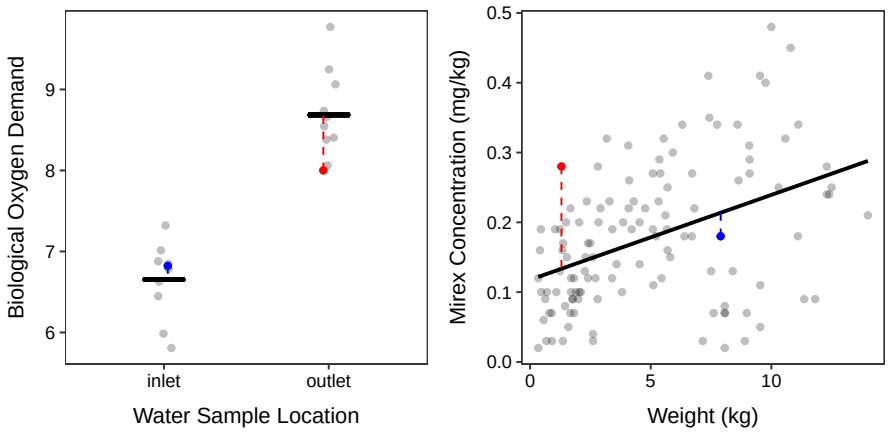


Figure 3.4: Same plots as previously but with a "large" residual shown in red and a "small" residual shown in blue.

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.

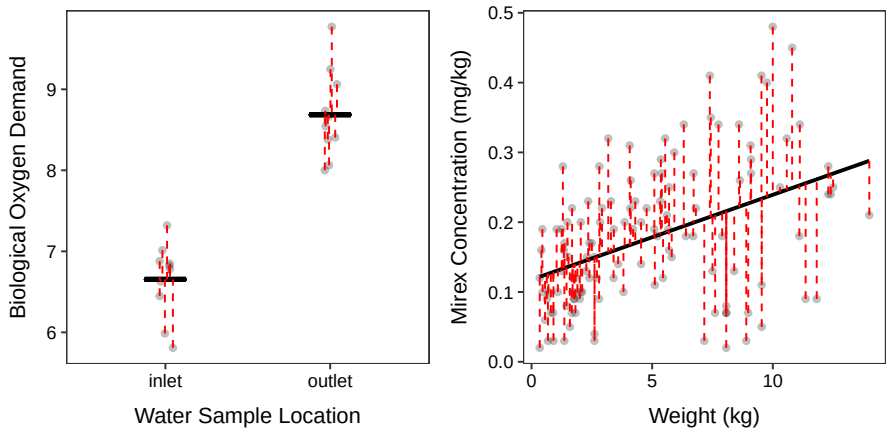


Figure 3.5: Same plots as previously but with all residuals shown.

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.

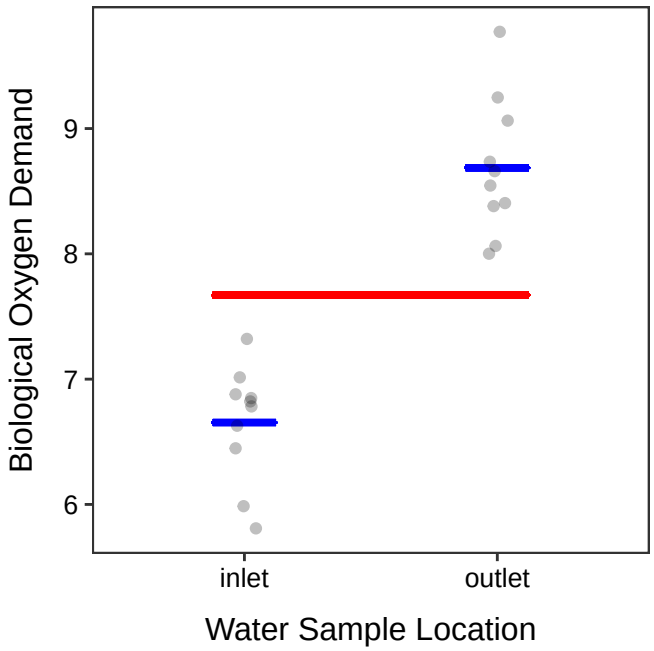


Figure 4.1: Biological oxygen demand versus sample location with group means shown as blue horizontal segments and the grand mean shown as a red horizontal segment.

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).

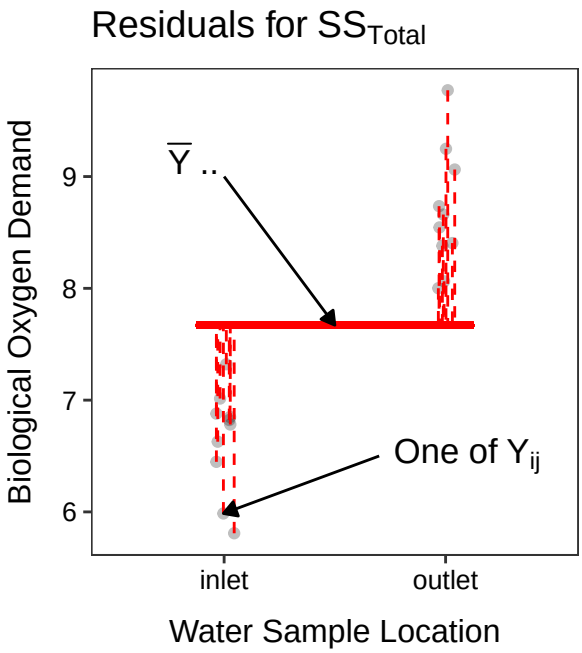


Figure 4.2: Biological oxygen demand versus sample location with the grand mean shown as a red horizontal segment. Residuals from the grand mean are shown by red vertical dashed lines. The sum of these residuals is SS_{Total}.

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).

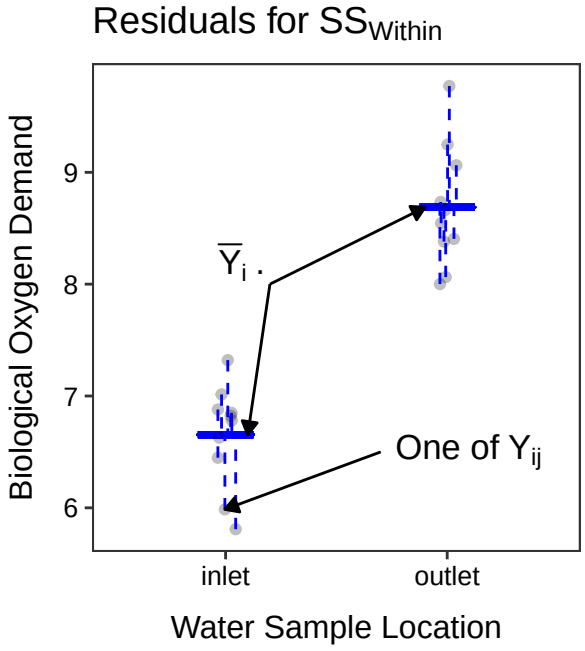


Figure 4.3: Biological oxygen demand versus sample location with the group means shown as blue horizontal segments. Residuals from the group means are shown by blue vertical dashed lines. The sum of these residuals is SS_{Within}.

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

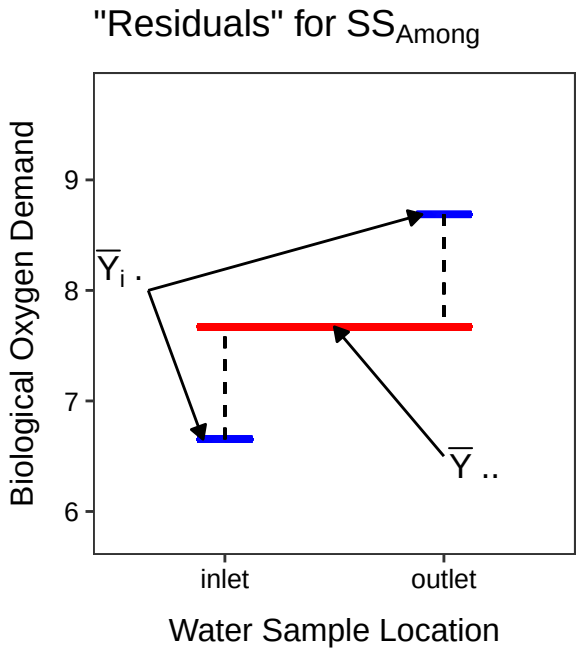
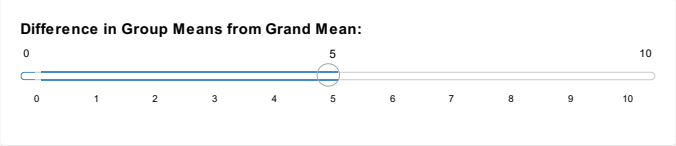


Figure 4.4: Mean biological oxygen demand versus sample location with the grand mean shown as a red horizontal segment and the group means shown as blue horizontal segments. Residuals between the group means and the grand mean are shown by black vertical dashed lines. The sum of these residuals scaled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

Partitition Total SS

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).



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.,

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

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

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

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

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

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

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.

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

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.,

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

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.

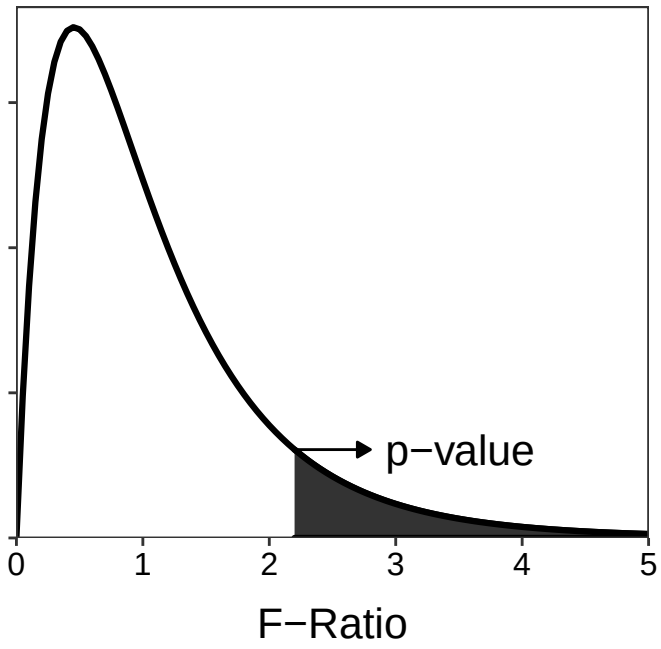
The question now becomes "when is the F-ratio considered large enough to reject the simple model and conclude that the full model is significantly better?" This question can by comparing the F-ratio test statistic to an F-distribution.

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.

[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.



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 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.

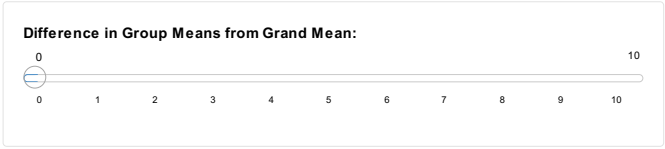


Table 4.2: An ANOVA table for biological oxygen demand measurements at two locations of the aquaculture facility. Note that the "Total" row is not shown.

	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 ANOVA 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.

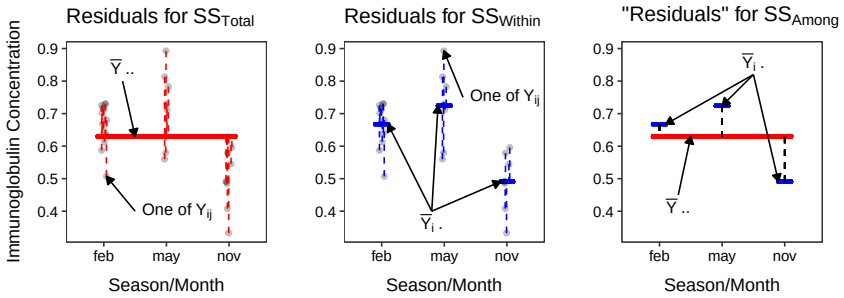


Figure 5.1: Immunoglobulin concentrations versus season (or month) of capture for New Zealand opposums. The grand mean is shown by a red horizontal segment, group means are shown by blue horizontal segments, residuals from the grand mean are red vertical dashed lines, residuals from the groups means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

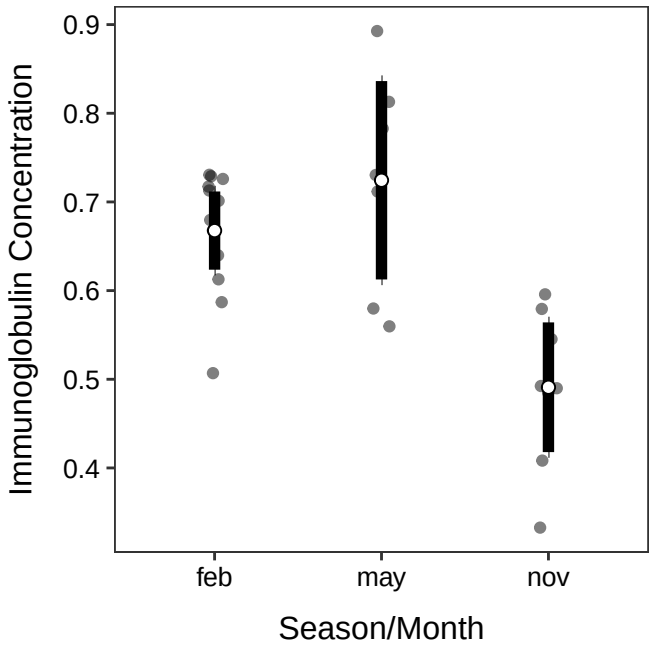
```
lm1 <- lm(imm~season,data=opp)
anova(lm1)
```

```
#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

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**[1] because it relates to only one comparison of a pair of means.

[1] This is sometimes called an comparison-individual-, or test-wise error.

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**[2] 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!!

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.[3] 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.

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.[4] 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

[2] This is sometimes called a family-wise error.
[3] Using =0.05
[4] See here for a short list of methods.

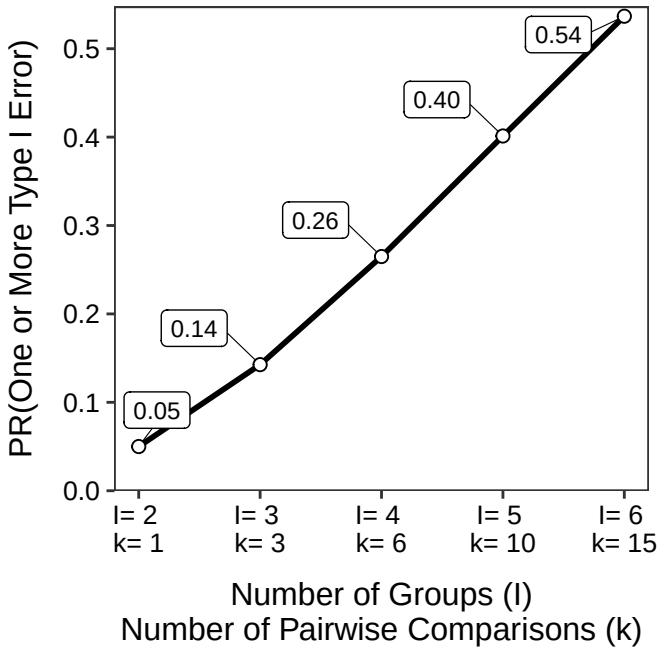


Figure 6.1: Relationship between the number of groups (I) in an analysis, the number of pairs of means that would need to be tested (k), and the probability of making one or more Type I errors in all comparisons. Note that alpha=0.05.

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 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

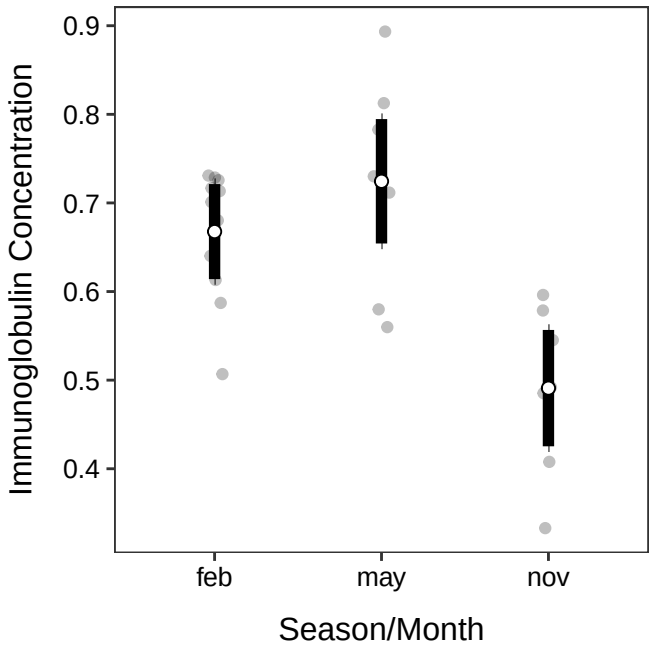
section tests if the group population mean is different than 0. These tests are not often of interest and can largely be ignored.

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()
```


Immunoglobulin Concentration
0.9
0.8
0.7
0.6
0.5
0.4
feb
may
nov
Season/Month