MODULE 1

FOUNDATIONS

Module Objectives:

1. Understand how models are related to hypothesis tests.
2. Understand the concept underlying comparing a simple and full model.
3. Understand how MS_{Total}, MS_{Within}, and MS_{Among} are computed.
4. Understand what MS_{Total}, MS_{Within}, and MS_{Among} measure or represent.
5. Understand how the F test statistic is used to make a decision regarding two statistical models.
6. Understand how the decision about two statistical models relates to a decision about hypotheses.
7. Understand how to read an ANOVA table and identify relationships within the table (e.g., partitioning).

THREE TYPES OF LINEAR MODELS, characterized by the type of response variable and type(s) of explanatory variables, will be considered in this book. The first major type is the "ANalysis Of VAriance" or ANOVA models. ANOVA models are characterized by a quantitative response variable and categorical(factor or grouping) explanatory variables (Table 1.1). One-way ANOVAs have only one categorical explanatory variable, whereas two-way ANOVAs have two categorical explanatory variables.

Table 1.1. Selected linear models categorized by the types of response and explanatory variables.

Explanatory Variable(s)	Response Variable: Quantitative	Response Variable: Categorical
Categorical	One-Way ANOVA (Module 2) Two-Way ANOVA (Module 3)	Chi-Square
Quantitative	Linear Regression (Module 4)	Logistic Regression (Module 6)
Both	Indicator Variable Regression (Module 5)	

The second major type of linear model to be considered is linear regression, which are characterized by quantitative response and explanatory variables (Table 1.1). Simple linear regression (SLR) occurs when only one explanatory variable is considered, whereas multiple linear regression considers multiple explanatory variables. Multiple quantitative explanatory variables will not be explored here. However, the third major type of linear model is indicator variable regression (IVR), which is a multiple regression with one quantitative and one or more categorical (factor or grouping) explanatory variables (Table 1.1). The ANalysis of COVAriance (ANCOVA) model is a special case of indicator variable regression.

The fourth major type of linear model to be considered is logistic regression, where the response variable is categorical and the explanatory variables are generally quantitative (Table 1.1).[1]

The remainder of this module is devoted to reviewing basic concepts from your introductory statistics course to develop a common framework that can be used to analyze all four types of these linear models. Thus, this module serves as the theoretical foundation for Modules 2-6.

1.1 Two-Sample t-Test 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 are covered in most introductory statistics courses.[2] Specifically, the null hypothesis is $H_0 : \mu_1 = \mu_2$, where the subscripts represent the two populations. This hypothesis is tested with the t test statistic,[3]

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{1.1.1}$$

where the pooled sample variance, s_p^2, is a measure of the common variance found within each population

[1] Though categorical explanatory variables may be considered in a manner similar to indicator variable regression models.
[2] Thus, it is assumed that you have a working knowledge of a two-sample t-test.
[3] Under the assumptions of independent and normally distributed "errors" and equal variances between the two groups.

and is computed with

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

Methods for analyzing a two-sample t-test in R are briefly summarized in the next three subsections.

1.1.1 Data Format

The data for a 2-sample t-test must be entered in stacked format. In stacked format, the measurements are in one column (or vector) and a label for the populations is in another column. Each row of the data.frame corresponds to the measurement and population of a single individual. The data for the example below are the biological oxygen demands at the inlet and outlet to an aquaculture facilty. These data are read from the **BOD.csv** (view, download, meta) file below[4] and several rows are displayed with `headtail()`.

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

```
> headtail(aqua)   # 1st and last 3 rows, not whole data frame
     BOD    src
1  6.782  inlet
2  5.809  inlet
3  6.849  inlet
18 8.545 outlet
19 8.063 outlet
20 8.001 outlet
```

Δ **Stacked Data**: Data where the quantitative measurements of two 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 the preferred format for two-sample data because each column corresponds to a variable and each row corresponds to only one individual.**

1.1.2 Levene's Test

Before conducting a 2-sample t-test, the assumption of equal variances must be tested with Levene's test. The first argument to `levenesTest()` is a model formula of the type `response~factor`, where `response` represents the variable that contains the quantitative measurements and `factor` represents the variable that contains the categorical groups. In addition, the data.frame in which the `response` and `factor` variables are found is given to `data=`. The very large p-value (=0.591) in the BOD example indicates that the variances can be considered to be equal.

[4] These data can be downloaded from the class webpage by right-clicking on "download" and saving to your computer. The working directory should then be set to where this file is located on your computer. The data are then read into R with `read.csv()`. It is assumed that you remember this procedure from your introduction to R.

```
> levenesTest(BOD~src,data=aqua)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.2989 0.5913
      18
```

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

1.1.3 The Test

A 2-sample t-test is constructed in R with `t.test()`, with the exact same `response~factor` formula used in `levenesTest()` as the first argument and the corresponding data.frame in `data=`. Additionally, the following arguments may be specified:

- `var.equal`: A logical that indicates whether the two population variances should be considered equal or not. If `var.equal=TRUE`, then the pooled sample variance is calculated and used in the standard error. The default value is to NOT assume unequal variances; thus, `var.equal=TRUE` must typically be set.
- `alternative`: A character string that indicates whether the alternative hypothesis is `"two.sided"` (i.e., the "not equals" situation), `"greater"`, or `"less"`. The default is `"two.sided"`.
- `conf.level`: The proportional level of confidence to be used when constructing the confidence interval for $\mu_1 - \mu_2$. The default is `0.95`.

◇ **The `var.equal=TRUE` argument must be used in `t.test()` to assume equal variances. This is NOT the default setting in R.**

Finally, the order of levels of the grouping factor is alphabetical by default. This results in the mean for the alphabetically-second population being subtracted from the mean of the alphabetically-first populations. In this example, the default 2-sample t-test would subtract the *outlet* mean from the *inlet* mean. If this is opposite of the way the hypotheses were set up, then either (i) the order of the groups in the hypotheses must be reversed (i.e., $H_A : \mu_{out} - \mu_{in} > 0$ is the same as $H_A : \mu_{in} - \mu_{out} < 0$) or (ii) the order of the levels of the grouping factor must be explicitly set with `factor()`. For example, *outlet* is forced to be the first level in (`src`) below.

```
> aqua$fsrc <- factor(aqua$src,levels=c("outlet","inlet"))
> levels(aqua$fsrc)
[1] "outlet" "inlet"
```

When using `factor()`, be very careful to type the level names exactly as they appear in the original variable. For example, if `levels=c("Outer","Inlet")` had been used, then the *src* variable would contain only `<NA>` values, which is Rs way of saying "not available." In other words, a variable with nothing in it would have been created.

◇ **When assigning the order of the levels of a factor variable, take care to use level names exactly as they appeared in the original variable.**

The `factor()` function is also used to tell R to treat a particular variable as a factor variable. For example, suppose that the researcher had entered the numbers 1 and 2 for the *src* variable. By default, this variable would be treated as numeric rather than as a factor with levels. Including the original variable name in `factor()` and assigning it to a new variable will force R to treat the new variable as a factor.

◇ **Factor variables that were labeled with numbers must be specifically told to be treated as a factor variable with `factor()`.**

The 2-sample t-test of the BOD data, using the new level ordering and assuming equal variances (per the Levene's Test results above), is computed below. From these results it is seen that the difference between the sample means at the outlet and the inlet is 8.687-6.654 = 2.034, the test statistic is 8.994 with 18 df, and the p-value is < 0.0005. Thus, there does appear to be a significant difference between the population mean BOD of water at the inlet and outlet to the aquaculture facility. Furthermore, one is 95% confident that the mean BOD at the outlet is between 1.558 and 2.509 greater than the mean BOD at the inlet.

```
> t.test(BOD~fsrc,data=aqua,var.equal=TRUE)   # uses default alt= & conf.level=
 Two Sample t-test with BOD by fsrc
t = 8.994, df = 18, p-value = 4.449e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.558489 2.508511
sample estimates:
mean in group outlet  mean in group inlet
              8.6873               6.6538
```

1.2 Models

Many hypothesis tests, including the two-sample t-test, can be cast in a framework of competing models. This conceptualization for the two-sample t-test in this secton will serve as the conceptual foundation for all other linear models in this course.

The null and alternative hypotheses are each related to a model. The null hypothesis corresponds to the *simple* model, whereas the alternative hypothesis corresponds to the *full* model.[5] In the two-sample t-test, the null hypothesis represents the situation of no difference between population means. Thus, a single mean, μ, would represent both populations. In contrast, the alternative hypothesis implies that a difference exists between the two population means. Thus, each population must be represented by a separate mean (e.g., μ_1 and μ_2). With this, the simple and full models for a two-sample t-test are

$$simple : \mu_i = \mu$$
$$full : \mu_i = \mu_i$$

[5]Note that the only simple and full model used in this foundational development are sometimes called the *ultimate simple* and *ultimate full* models because there can be no model simpler than just using a grand mean and there can be no model more complicated then using a separate mean for each group. There will be other "simple" and "full" models throughout this course, but when the words "ultimate simple" and "ultimate full" model are used then reference is made to these two specific models.

where i represents the ith group (i.e, would be replaced with "1" or "2"), μ_i represents the population mean of the ith group, and μ (with no subscript) represents the *grand* mean for both groups combined. Thus, the simple model says that there is one mean – the grand mean, μ – that adequately represents each group; whereas the full model says that each group is represented by a separate mean. These models are visually represented in Figure 1.1. The simple model is called "simple" because it has fewer parameters (i.e., one mean) than the full model (i.e., as many means as groups).

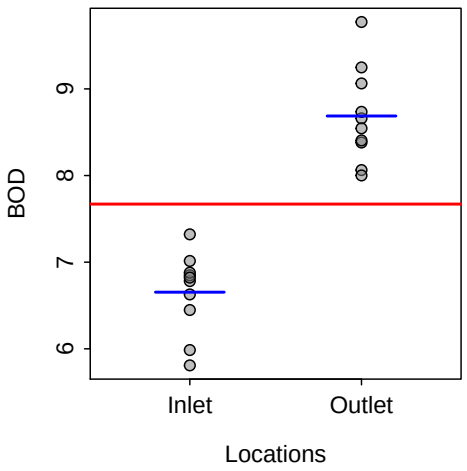

Figure 1.1. Biological oxygen demand (BOD) measurements at two locations – an inlet and an outlet. The red horizontal line represents the simple model of a grand mean for both groups. The two blue horizontal lines represent the full model of separate means for each group.

◇ **The null hypothesis model always has fewer parameters and is thus called the "simple" model. The alternative hypothesis model always has more parameters and is thus called the "full" model.**

◇ **The simple model in a two-sample t-test represents a flat line at the grand mean of the response variable. The full model in a two-sample t-test represents a separate line at the means of the response variable for both groups.**

1.3 Sum-of-Squares

When comparing two competing models in statistics, an attempt is made to determine if the simple model fits the data "as well as" the full model. It is important to note that the full model will always fit the data, at least somewhat, better. Therefore, it must be determined whether the full model fits the data enough better to warrant the use of the extra parameter(s). In other words, if the full model does not fit significantly better, then the added complexity of the full model is not warranted. So, a measure that allows the comparison of how well two models fit the data relative to how many parameters are in each model is needed. An initial step in the computation of this measure is developed in this section.

The general lack-of-fit of a model is measured by computing the residuals of the individual observations using

predicted values derived from the model in question. Models with relatively small residuals, in some total sense, are considered "good" models. However, because the residuals will always sum to zero, the overall measure of model lack-of-fit is found by summing the square of the residuals.[6] Very generally then, the sum of squared residuals will have this form

$$\text{Sum of Squared Residuals} = \sum_{i=1}^{n} (Observed - Predicted)^2 \tag{1.3.1}$$

where n is the number of individuals.

◇ **A residual is always computed as the difference between an observed value of the response variable and a value of the response variable predicted by using a model.**

◇ **An overall measure of the lack-of-fit of a model is the sum of the squared residuals computed using that model.**

Some notation must be defined before this general formula can be made more specific. Let Y represent a generic response variable.[7] Furthermore, let Y_{ij} be the measurement of the response variable for the jth individual in the ith group. Thus, j is an index for individuals within a group and i is an index for groups. For example, Y_{23}, is the measurement of the response variable on the third individual in the second group. The number of individuals in group i is depicted by n_i. The sample mean of the individuals in the ith group is $\bar{Y}_{i\cdot}$ and is computed separately for each group as[8]

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

The $\bar{Y}_{i\cdot}$ are called "group means" or, sometimes, "treatment means." Finally, $\bar{Y}_{\cdot\cdot}$ represents the grand or overall mean of all individuals in the study and is computed as

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

where I represents the total number of groups (= 2 for a two-sample t-test) and, again, n represents all individuals in the study and is thus

$$n = \sum_{i=1}^{I} n_i$$

With these symbols and Equation (1.3.1), the overall lack-of-fit for the simple model is measured by calculating the residuals with predictions made from the grand mean[9] of the response variable (i.e., $\bar{Y}_{\cdot\cdot}$). Thus, the overall lack-of-fit for the simple model is measured by

[6] It is assumed that you are generally familiar with the concept of "minimizing sum-of-squares to identify the best model" by being exposed to the basics of linear regression in your introductory statistics course

[7] Note that this is a little different then what was likely done in your introductory statistics course where X usually represented a generic variable. It is common in advanced statistics to call Y the generic response variable as the response variable is usually plotted on the y-axis.

[8] Note that it is common practice to put a dot where the subscript that was summed across would be. In this example, the individuals within a group were summed, or summing was across the j index, thus the j is replaced with a dot.

[9] Recall that the simple model states that one mean, the grand mean, represents all treatment groups.

$$SS_{Total} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{..}\right)^2 \tag{1.3.2}$$

The sum-of-square residuals, or sum-of-squares (SS) for short, for this particular simple model (i.e., using $\bar{Y}_{..}$) is called the total SS, or SS_{Total}, because it is the basis of the measure of the "total" variability in the response variable.[10]

◇ **The SS_{Total} measures the lack-of-fit of the simplest model using a single common mean to represent each group.**

The overall lack-of-fit for the full model is measured by calculating the residuals with predictions computed using separate group means (i.e., $\bar{Y}_{i.}$).[11] Thus, the overall lack-of-fit of the full model is measured by

$$SS_{Within} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i.}\right)^2 \tag{1.3.3}$$

The sum-of-squares for this full model (i.e., using $\bar{Y}_{i.}$) is called SS_{Within} because it measures the lack-of-fit of individuals **within** each group from the mean for that group. In other words, it measures the SS of individuals **within** each group and then combines this measure for all groups to form one overall SS of individuals within groups.[12]

◇ **The SS_{Within} measures the lack-of-fit of the fullest model using a separate mean to represent each group.**

It is very important to understand the critical difference between Equation (1.3.2) and Equation (1.3.3) (i.e., using $\bar{Y}_{..}$ versus using $\bar{Y}_{i.}$). A close examination of Figure 1.2 shows that SS_{Total} is the sum of the squared residuals computed from each individual (i.e., dot) to the long horizontal line at the grand mean on the left graph. SS_{Within}, on the other hand, is the sum of the squared residuals computed from each individual within a group to the short horizontal lines at each group mean in the right graph. These two sums-of-squares measure two completely different SS; i.e., lacks-of-fit for two different models for the data! A residual is a measure of how "far off" a particular model is from a particular point. The sum of the square of these residuals (SS) is a measure of how "far off" a particular model is from all of the points in a data set. No model can perfectly represent all individuals; the SS is a measure of this imperfectness, or the "lack-of-fit", by the model. Specifically, SS_{Total} measures the lack-of-fit of the simple model and SS_{Within} measures the lack-of-fit of the full model.

◇ **Sums-of-squares measure the lack-of-fit to the data by a particular model.**

[10] It will be shown in Section 1.4 that dividing Equation (1.3.2) by $n-1$ will result in s^2_Y - the sample variance of the response variable Y.

[11] Recall that the full model states that a separate mean is used for each group.

[12] Recall that a major assumption of a two-sample t-test is that the variance is the same for each group. If the variance is the same for each group then there is really only one variance to estimate. Thus, the estimates from separate groups are pooled together to form this single estimate. You will see in Section 1.4 that dividing SS_{Within} by $n-I$ provides this pooled estimate of the common variance.

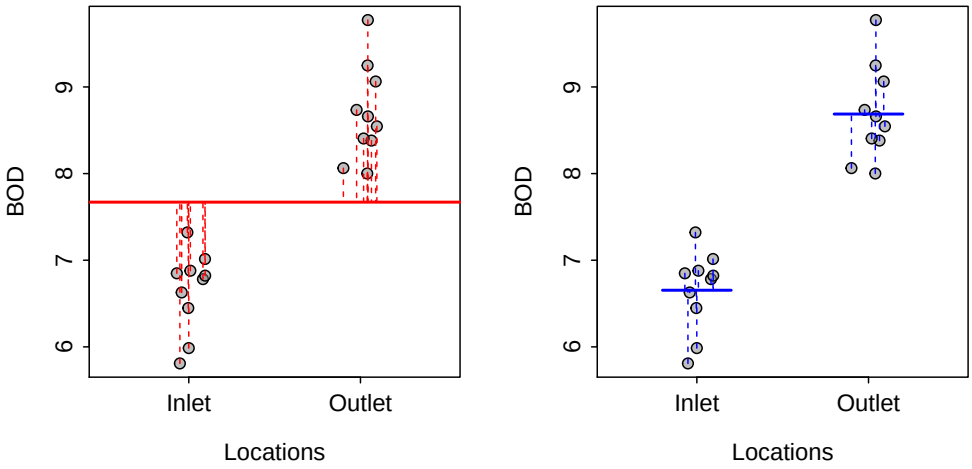

Figure 1.2. Biological oxygen demand (BOD) measurements at two locations – an inlet and an outlet – with the simple model (Left) and full model (Right) shown. In addition, residuals for each model are shown on the respective graphs. Note that the points were horizontally "jittered" so that each residual could be seen.

◇ SS_{Total} **measures the lack-of-fit of the simple model.** SS_{Within} **measures the lack-of-fit of the full model.**

Partitioning SS

It can be shown algebraically that SS_{Total} can be separated into two parts.

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{\cdot\cdot}\right)^2=\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i\cdot}\right)^2+\sum_{i=1}^{I}n_i\left(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot}\right)^2 \tag{1.3.4}$$

The first term on the right-hand side is SS_{Within}. Thus,

$$SS_{Total}=SS_{Within}+\sum_{i=1}^{I}n_i\left(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot}\right)^2$$

The remaining term on the right-hand side is called SS_{Among}. Thus,

$$SS_{Among}=\sum_{i=1}^{I}n_i\left(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot}\right)^2 \tag{1.3.5}$$

and SS_{Total} thus partitions into two generic parts,

$$SS_{Total}=SS_{Within}+SS_{Among} \tag{1.3.6}$$

SS_{Among} is a critically important statistic in the comparison of two statistical models because it represents the difference in lack-of-fit for the simple model and the lack-of-fit for the full model; i.e., $SS_{Among}=SS_{Total}-SS_{Within}$. In fact, SS_{Among} represents the improvement in lack-of-fit that is achieved by using the full model as compared to the simple model. Thus, SS_{Among}, represents a measure of how much "better" the full model represents the data as compared to the simple model. In this example, SS_{Among} is a measure of how much the fit is improved by using separate means for each group rather than a common grand mean.

◇ **The lack-of-fit from using the simple model (i.e., SS_{Total}) can be partitioned into two parts – the lack-of-fit from using the full model and the improvement in lack-of-fit that was gained by using the full model over the simple model.**

◇ **The within SS (SS_{Within}) is the measure of the lack-of-fit when using the full model.**

◇ **The among SS (SS_{Among}) is the measure of the improvement in lack-of-fit from using the full over the simple model.**

Visually, SS_{Among} is the difference between the total vertical spread in individuals regardless of the group (i.e., SS_{Total}) and the average vertical distance among individuals within the groups (i.e., SS_{Within}). Alternatively, the SS_{Among} can be visualized as the total vertical spread[13] among the horizontal lines representing the different group means (Figure 1.3).

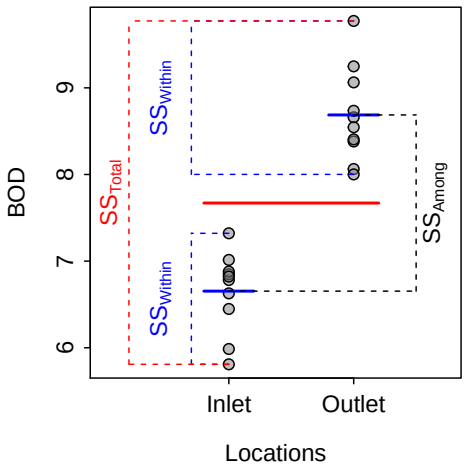

Figure 1.3. Biological oxygen demand (BOD) measurements at two locations – an inlet and an outlet – with the simple model (solid red line) and full model (solid blue line) shown. In addition, representations of SS_{Total}, SS_{Within}, and SS_{Among} are shown.

It seems intuitive that SS_{Among} can be used to judge which model fits the data "better" because it seems reasonable that if SS_{Among} is greater than 0 then the full model is "better" than the simple model. However, SS_{Among} is always greater than 0 because the full model is always at least somewhat "better" then the simple model. Furthermore, SS_{Among} (and all other SS) is a statistic that is subject to sampling variability. The issues of model complexity (i.e., the number of parameters) and sampling variability are addressed in the following two sections.

◇ **The SS_{Among} measures "how much better" a full model fits the data as compared to a simple model. However, SS_{Among} cannot be effectively used to compare the full and simple models because it is always greater than 0 and is subject to sampling variability.**

[13] Necessarily re-scaled to represent the sample size within each group.

1.4 Mean Squares

The SS are not true measures of variability – they must be divided by their corresponding degrees-of-freedom (df) to be a variance. When a SS is divided by the corresponding df it is called a mean-square. Thus, mean-squares are true variances.

◇ **Mean-squares are equal to SS divided by df. Mean-squares are variances.**

The variance about the simple model is thus measured by

$$MS_{Total} = \frac{SS_{Total}}{df_{Total}} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{\cdot\cdot}\right)^2}{n-1} = s_Y^2 \tag{1.4.1}$$

and the variance about the full model is thus measured by

$$MS_{Within} = \frac{SS_{Within}}{df_{Within}} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2}{n-I} = s_p^2 \tag{1.4.2}$$

From Equation (1.4.1) and Equation (1.4.2), it is evident that the total df is equal to $n-1$ and the within df is equal to $n-I$. Furthermore, the df partition in the same way that the SS partition. Thus,

$$\begin{aligned} df_{Total} &= df_{Within} + df_{Among} \\ n-1 &= n-I+df_{Among} \end{aligned}$$

where, by subtraction, the df_{Among} is then equal to $I-1$. The df_{Among} represent the difference in number of parameters between the simple and full model.

◇ **The among df (df_{Among}) is equal to the difference in number of parameters between the full and simple models.**

Thus, the calculation of MS_{Among},

$$MS_{Among} = \frac{SS_{Among}}{df_{Among}} = \frac{\sum_{i=1}^{I} n_i\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2}{I-1} \tag{1.4.3}$$

rectifies the first issue with SS_{Among} identified above; i.e., the number of parameters in the models must be taken into account in the calculations. Furthermore, MS_{Among} can be interpreted as the variance around the simple model that is explained by the full model.

◇ **MS_{Among} represent the variability in the simple model that is explained by using the full model.**

Sources of Variability in Variances

The MS_{Total} is a measure of the total natural variability in the response variable – i.e., the variability of the response variable among individuals without regard to any other (explanatory) variable(s). So, MS_{Total} measures the maximum variability in the response variable. For example, MS_{Total} is a measure of the total and maximum amount of variability evident in BOD among individuals.[14]

◇ **The MS_{Total} measures the variability in the simplest model, which is the grand mean of the response variable. Thus, MS_{Total} measures the maximum total variability in the response variable.**

The total variability in the response variable may be a result of differences among individuals (i.e., natural variability) or differences among sample means. Differences among sample means may be random (i.e., sampling variability) or real (i.e., means differ among populations). These sources of variability are measured by MS_{Within} and MS_{Among}. MS_{Within} is a measure of the natural variability within each group and, thus, is unaffected by different means between the groups. MS_{Among}, on the other hand, is computed from the differences in group sample means and, thus, is a measure of the variability between or among group means.[15] MS_{Among} is affected by both sampling variability and any real differences between the population means. In the next section, sampling variability is accounted for such that a measure of the real differences among groups will remain. This real difference forms the basis of a hypothesis test for determining which model best "fits" the data.

◇ **The total variability in the response variable may result from three sources – (1) natural variability, (2) sampling variability, or (3) real differences in population means.**

1.5 F-test

1.5.1 Overall

The issue with sampling variability is rectified by comparing the variability explained by the full model to the variability unexplained by the full model. In other words, the explained variability is "scaled" by the unexplained variability (the F will be explained shortly).

$$F = \frac{MS_{Among}}{MS_{Within}} \tag{1.5.1}$$

If this ratio is "large," then a great deal more variability was explained than was unexplained by the full model 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 ratio in Equation (1.5.1) considered large enough to reject the simple model and conclude that the full model is significantly better?" This question can be answered by realizing that F computed in Equation (1.5.1) is a test statistic that follows an F distribution.

[14] This concept is illustrated by SS_{Total} in Figure 1.3

[15] Note that the word "between" is used for comparing two groups whereas the word "among" is used for comparing more that two groups. The word "among" is more general and will be used throughout these notes, regardless of how many groups are being compared.

An F-distribution occurs whenever the ratio of two variances is calculated. An F distribution (Figure 1.4) 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 MS_{Among}. The denominator df is equal to the df used in MS_{Within}. The p-value is always computed as the area under the F-distribution curve to the right of the observed F statistic (Figure 1.4).[16]

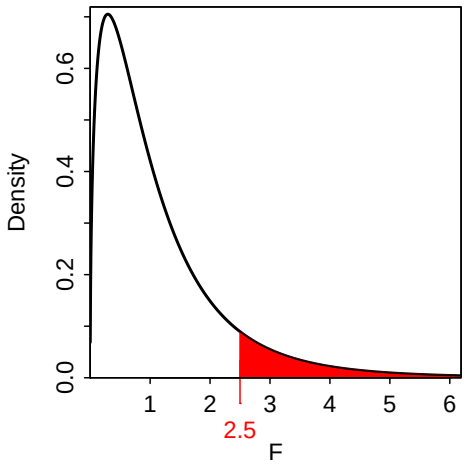

Figure 1.4. Example F-distribution showing the p-value if the observed F test statistic was 2.5.

◇ **An F test statistic that follows an F-distribution arises whenever the ratio of two variances is computed.**

◇ **The exact F-distribution is dictated by so-called numerator and denominator df.**

◇ **The p-value from an F-distribution is always computed as the area in the upper-tail.**

If the computed p-value is less than α, then it is concluded that the variability explained by the full model is significantly greater than the variability left unexplained by the full model. In other words, the null hypothesis is rejected in favor of the alternative hypothesis and it is concluded that the full model is significantly better than the simple model. In the case of a two-sample t-test, this means that the mean value of the response variable differs between the two populations.

◇ **A p-value computed from an F test statistic that is less than α indicates that the full model is significantly "better" then the simple model, even after taking into account the increased complexity of the full model and sampling variability.**

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

1.5.2 General

The F test shown in Equation (1.5.1) is specific to comparing the ultimate full and ultimate simple models. It will be shown in later modules that these are not the only two full and simple models that will be compared. Thus, a more general formula for the F-test statistic is

$$F = \frac{\frac{RSS_{Simple} - RSS_{Full}}{df_{Simple} - df_{Full}}}{\frac{RSS_{UltimateFull}}{df_{UltimateFull}}} = \frac{\frac{RSS_{Simple} - RSS_{Full}}{df_{Simple} - df_{Full}}}{RMS_{UltimateFull}} \tag{1.5.2}$$

where RSS is $SS_{residual}$ and RMS is $MS_{residual}$ from fitting the model. In the discussions of the previous section, the RSS was measured by SS_{Within}; the RSS notation is a bit more general.

1.6 ANOVA Table

The degrees-of-freedom (df), sum-of-squares (SS), mean-squares (MS), F test statistic (F), and corresponding p-value are summarized in an analysis of variance, or ANOVA, table (e.g., Table 1.2). The ANOVA table contains rows that correspond to the different sources of variability that were discussed above: among,[17] within,[18] and total. The df and SS are shown for each source of variability. The MS is shown for the within and among sources, but generally not for the total because the MS for within and among sources do not sum to the MS for total (the SS and df partition, but the MS do not!).

Table 1.2. Analysis of variance table for the BOD measurements at an inlet and outlet sources.

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
src        1 20.6756 20.6756  80.891 4.449e-08 ***
Residuals 18  4.6008  0.2556
Total     19 25.2764
```

The results in Table 1.2 indicate that H_0 should be rejected (i.e., F-test p-value < 0.0005). 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 the 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 Table 1.2 as an example, note the following items:

1. The variance within groups is equal to MS_{Within} (e.g., $MS_{Residuals} = 0.2556$ in this case). This is analogous to s_p^2 from the two-sample t-test (in fact, it is exactly equal to s_p^2, if there are only two groups, as is the case here).
2. The common variance about the mean (s_Y^2) is given by MS_{Total} (e.g., $= \frac{25.2764}{19} = 1.3303$ in this case).

[17] Labeled as the factor variable in most statistical software packages including R – that variable was called *src* in this example.
[18] Labeled as residuals in R and error in other statistical software packages.

1.7 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 μ, 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 μ because the sample mean is an unbiased estimator of μ. 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.

MS_{Within} and MS_{Among} are statistics just as $\bar{x}$ and s^2 are statistics. Thus, MS_{Within} and MS_{Among} are subject to sampling variability and have sampling distributions. It can be shown[19] that the center of the sampling distribution of MS_{Within} is σ^2 and the center of the sampling distribution of MS_{Among} is

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

Thus, MS_{Among} 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., $\mu_1 = \mu_2 = \cdots = \mu_I = \mu$ – then the second source of variability is equal to zero and MS_{Among} will equal MS_{Within}. As soon as the groups begin to differ, the second source of variability will be greater than 0 and MS_{Among} will be greater than MS_{Within}.

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 MS_{Within} and MS_{Among} 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 H_0 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 MS_{Within} is σ^2 but the center of the sampling distribution of MS_{Among} 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 1 and will get larger as the null hypothesis gets "more false." This analysis of sampling distribution theory illustrates once again that (1) MS_{Among} consists of multiple sources of variability and (2) "large" values of the F test-statistic indicate that the null hypothesis is incorrect.

1.8 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 1.1. The results of `lm()` should be assigned to an object so that specific results can be selectively extracted. For example, the ANOVA table results are extracted from the `lm()` object with `anova()`. In addition, coefficient results[20] 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.

[19] This derivation is beyond the scope of this book.
[20] The coefficient results will be discussed in more detail in Module 2.

```
> aqua.lm <- lm(BOD~src,data=aqua)
> anova(aqua.lm)
Analysis of Variance Table

Response: BOD
          Df  Sum Sq Mean Sq F value    Pr(>F)
src        1 20.6756 20.6756  80.891 4.449e-08
Residuals 18  4.6008  0.2556
> cbind(ests=coef(aqua.lm),confint(aqua.lm))
              ests    2.5 %   97.5 %
(Intercept) 6.6538 6.317917 6.989683
srcoutlet   2.0335 1.558489 2.508511
> summary(aqua.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.6538     0.1599  41.619  < 2e-16
srcoutlet     2.0335     0.2261   8.994 4.45e-08

Residual standard error: 0.5056 on 18 degrees of freedom
Multiple R-squared: 0.818,Adjusted R-squared: 0.8079
F-statistic: 80.89 on 1 and 18 DF,  p-value: 4.449e-08
```

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.

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.