13.2 - The ANOVA Table

For the sake of concreteness here, let's recall one of the analysis of variance tables from the previous page:

Source	DF	SS	MS	F	P
Factor	2	2510.5	1255.3	93.44	0.000
Error	12	161.2	13.4		
Total	14	2671.7			

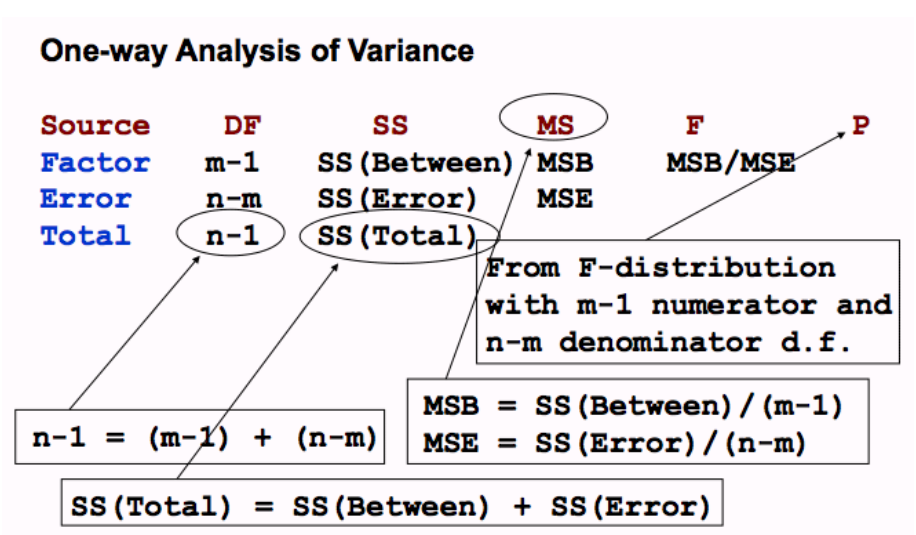
In working to digest what is all contained in an ANOVA table, let's start with the column headings:

1. **Source** means "the source of the variation in the data." As we'll soon see, the possible choices for a one-factor study, such as the learning study, are **Factor**, **Error**, and **Total**. The factor is the characteristic that defines the populations being compared. In the tire study, the factor is the brand of tire. In the learning study, the factor is the learning method.
2. **DF** means "the degrees of freedom in the source."
3. **SS** means "the sum of squares due to the source."
4. **MS** means "the mean sum of squares due to the source."
5. **F** means "the *F*-statistic."
6. **P** means "the *P*-value."

Now, let's consider the row headings:

1. **Factor** means "the variability due to the factor of interest." In the tire example on the previous page, the factor was the brand of the tire. In the learning example on the previous page, the factor was the method of learning.
   Sometimes, the factor is a treatment, and therefore the row heading is instead labeled as **Treatment**. And, sometimes the row heading is labeled as **Between** to make it clear that the row concerns the variation *between* the groups.
2. **Error** means "the variability within the groups" or "unexplained random error." Sometimes, the row heading is labeled as **Within** to make it clear that the row concerns the variation *within* the groups.
3. **Total** means "the total variation in the data from the grand mean" (that is, ignoring the factor of interest).

With the column headings and row headings now defined, let's take a look at the individual entries inside a general one-factor ANOVA table:



Yikes, that looks overwhelming! Let's work our way through it entry by entry to see if we can make it all clear. Let's start with the degrees of freedom (**DF**) column:

1. If there are *n* total data points collected, then there are ***n*–1** total degrees of freedom.
2. If there are *m* groups being compared, then there are ***m*–1** degrees of freedom associated with the factor of interest.
3. If there are *n* total data points collected and *m* groups being compared, then there are ***n*–*m*** error degrees of freedom.

Now, the sums of squares (**SS**) column:

1. As we'll soon formalize below, **SS(Between)** is the sum of squares between the group means and the grand mean. As the name suggests, it quantifies the variability between the groups of interest.
2. Again, as we'll formalize below, **SS(Error)** is the sum of squares between the data and the group means. It quantifies the variability within the groups of interest.
3. **SS(Total)** is the sum of squares between the n data points and the grand mean. As the name suggests, it quantifies the total variabilty in the observed data. We'll soon see that the total sum of squares, SS(Total), can be obtained by adding the between sum of squares, SS(Between), to the error sum of squares, SS(Error). That is:

$$\textbf{SS(Total) = SS(Between) + SS(Error)}$$

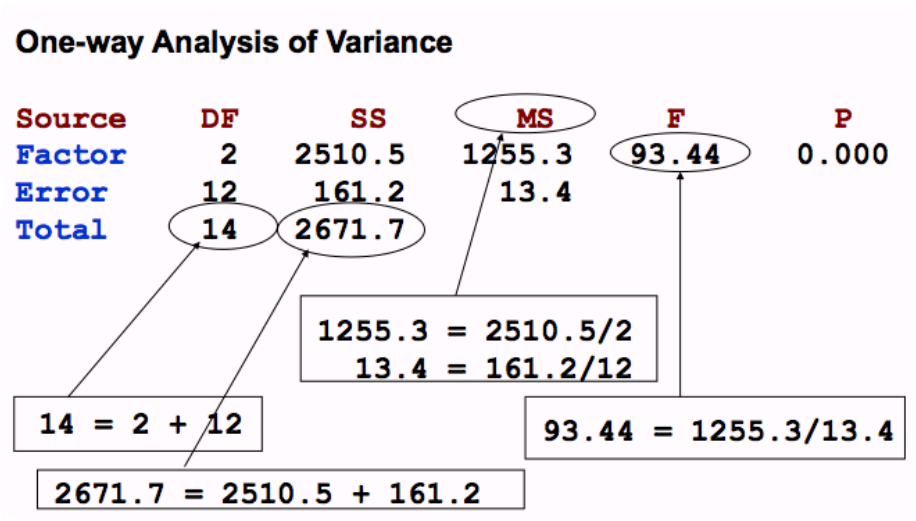
The mean squares (MS) column, as the name suggests, contains the "average" sum of squares for the Factor and the Error:

1. The Mean Sum of Squares between the groups, denoted **MSB**, is calculated by dividing the Sum of Squares between the groups by the between group degrees of freedom. That is, $\textbf{MSB = SS(Between)/(}m\textbf{–1)}$.
2. The Error Mean Sum of Squares, denoted **MSE**, is calculated by dividing the Sum of Squares within the groups by the error degrees of freedom. That is, $\textbf{MSE = SS(Error)/(}n\textbf{–}m\textbf{)}$.

The F column, not surprisingly, contains the F-statistic. Because we want to compare the "average" variability between the groups to the "average" variability within the groups, we take the ratio of the Between Mean Sum of Squares to the Error Mean Sum of Squares. That is, the F-statistic is calculated as $\boldsymbol{F}$ **= MSB/MSE**.

When, on the next page, we delve into the theory behind the analysis of variance method, we'll see that the F-statistic follows an F-distribution with m–1 numerator degrees of freedom and n–m denominator degrees of freedom. Therefore, we'll calculate the P-value, as it appears in the column labeled P, by comparing the F-statistic to an F-distribution with m–1 numerator degrees of freedom and n–m denominator degrees of freedom.

Now, having defined the individual entries of a general ANOVA table, let's revisit and, in the process, dissect the ANOVA table for the first learning study on the previous page, in which n = 15 students were subjected to one of m = 3 methods of learning:



1. Because n = 15, there are n–1 = 15–1 = 14 total degrees of freedom.
2. Because m = 3, there are m–1 = 3–1 = 2 degrees of freedom associated with the factor.
3. The degrees of freedom add up, so we can get the error degrees of freedom by subtracting the degrees of freedom associated with the factor from the total degrees of freedom. That is, the error degrees of freedom is 14–2 = 12. Alternatively, we can calculate the error degrees of freedom directly from n–m = 15–3=12.
4. We'll learn how to calculate the sum of squares in a minute. For now, take note that the total sum of squares, SS(Total), can be obtained by adding the between sum of squares, SS(Between), to the error sum of squares, SS(Error). That is:

$$2671.7 = 2510.5 + 161.2$$

5. MSB is SS(Between) divided by the between group degrees of freedom. That is, 1255.3 = 2510.5 ÷ 2.
6. MSE is SS(Error) divided by the error degrees of freedom. That is, 13.4 = 161.2 ÷ 12.
7. The F-statistic is the ratio of MSB to MSE. That is, F = 1255.3 ÷ 13.4 = 93.44.
8. The P-value is $P(F(2,12) \geq 93.44) < 0.001$.

Okay, we slowly, but surely, keep on adding bit by bit to our knowledge of an analysis of variance table. Let's now work a bit on the sums of squares.

The Sums of Squares

In essence, we now know that we want to break down the TOTAL variation in the data into two components:

1. a component that is due to the TREATMENT (or FACTOR), and
2. a component that is due to just RANDOM ERROR.

Let's see what kind of formulas we can come up with for quantifying these components. But first, as always, we need to define some notation. Let's represent our data, the group means, and the grand mean as follows:

Group	Data				Means
1	X_{11}	X_{12}	$\cdots$	X_{1n_1}	$\overline{X}_{1\cdot}$
2	X_{21}	X_{22}	$\cdots$	X_{2n_2}	$\overline{X}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	X_{m1}	X_{m2}	$\cdots$	X_{mn_m}	$\overline{X}_{m\cdot}$
		Grand Mean			$\overline{X}_{\cdot\cdot}$

That is, we'll let:

1. *m* denote the number of groups being compared
2. $\boldsymbol{X_{ij}}$ denote the $\boldsymbol{j_{th}}$ observation in the $\boldsymbol{i_{th}}$ group, where $\boldsymbol{i = 1, 2, \ldots, m}$ and $\boldsymbol{j = 1, 2, \ldots, n_i}$. Important thing to note here... note that *j* goes from 1 to $\boldsymbol{n_i}$, not to $\boldsymbol{n}$. That is, the number of the data points in a group depends on the group *i*. That means that the number of data points in each group need not be the same. We could have 5 measurements in one group, and 6 measurements in another.
3. $\boldsymbol{\bar{X}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}}$ denote the sample mean of the observed data for group *i*, where $\boldsymbol{i = 1, 2, \ldots, m}$
4. $\boldsymbol{\bar{X}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}}$ denote the grand mean of all *n* data observed data points

Okay, with the notation now defined, let's first consider the **total sum of squares**, which we'll denote here as ***SS*(*TO*)**. Because we want the total sum of squares to quantify the variation in the data regardless of its source, it makes sense that *SS*(*TO*) would be the sum of the squared distances of the observations $\boldsymbol{X_{ij}}$ to the grand mean $\boldsymbol{\bar{X}_{\cdot\cdot}}$. That is:

$$\boldsymbol{SS(TO) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\cdot\cdot})^2}$$

With just a little bit of algebraic work, the total sum of squares can be alternatively calculated as:

$$\boldsymbol{SS(TO) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}^2 - n\bar{X}_{\cdot\cdot}^2}$$

Can you do the algebra?

Now, let's consider the **treatment sum of squares**, which we'll denote ***SS*(*T*)**. Because we want the treatment sum of squares to quantify the variation between the treatment groups, it makes sense that *SS*(*T*) would be the sum of the squared distances of the treatment means $\boldsymbol{\bar{X}_{i\cdot}}$ to the grand mean $\boldsymbol{\bar{X}_{\cdot\cdot}}$. That is:

$$\boldsymbol{SS(T) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2}$$

Again, with just a little bit of algebraic work, the treatment sum of squares can be alternatively calculated as:

$$\boldsymbol{SS(T) = \sum_{i=1}^{m} n_i\bar{X}_{i\cdot}^2 - n\bar{X}_{\cdot\cdot}^2}$$

Can you do the algebra?

Finally, let's consider the **error sum of squares**, which we'll denote ***SS*(*E*)**. Because we want the error sum of squares to quantify the variation in the data, not otherwise explained by the treatment, it makes sense that *SS*(*E*) would be the sum of the squared distances of the observations $\boldsymbol{X_{ij}}$ to the treatment means $\boldsymbol{\bar{X}_{i\cdot}}$. That is:

$$\boldsymbol{SS(E) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2}$$

As we'll see in just one short minute why, the easiest way to calculate the error sum of squares is by subtracting the treatment sum of squares from the total sum of squares. That is:

$$\boldsymbol{SS(E) = SS(TO) - SS(T)}$$

Okay, so now do you remember that part about wanting to break down the total variation *SS*(*TO*) into a component due to the treatment *SS*(*T*) and a component due to random error *SS*(*E*)? Well, some simple algebra leads us to this:

$$\boldsymbol{SS(TO) = SS(T) + SS(E)}$$

and hence why the simple way of calculating the error of sum of squares. At any rate, here's the simple algebra:

> ## Proof
>
> Well, okay, so the proof does involve a little trick of adding 0 in a special way to the total sum of squares:

$$SS(TO)=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\left(X_{ij}-\bar{X}_{i\cdot}\right)+\left(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot}\right)\right)^2$$

Then, squaring the term in parentheses, as well as distributing the summation signs, we get:

$$SS(TO)=\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2+2\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})+\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})^2$$

Now, it's just a matter of recognizing each of the terms:

$$SS(TO)=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{i\cdot}\right)^2+2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{i\cdot}\right)\left(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot}\right)+\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot}\right)^2$$

That is, we've shown that:

$$SS(TO)=SS(T)+SS(E)$$

as was to be proved.

« Previous

Next »