Analysis of Variance (ANOVA) – Single Factor

In the past, when we needed to compare population means, we were restricted to determining the existence of a significant difference using only two data sets. Let's suppose that you, the statistician, needs to conduct an experiment involving several population means. To determine an existence of a significant difference between any of the means, we could compare each possible pair of means individually. However, this can become very tedious, and multiple *t*-tests increase the probability of making Type I errors. For example, if we had to compare 15 population means, then we would need to perform 105 tests in order to compare every pair of means possible. What we need is a test that can compare all 15 of the population means at once.

Let's suppose that you want to test the effect of sleep deprivation on Algebra test scores. You begin by sending an email out to all of the students taking MAT 101. From this, you get 20 volunteers to take part in the experiment. You create four sleep time categories: 2 hours, 4 hours, 6 hours, and 8 hours. The number of study time hours and the Algebra test will be the same for each student. You randomly choose five students from the 20 for each category. That night, they sleep their set number of hours and return the next day to take the test. The scores for each student are recorded.

Notice that the students were selected randomly for each sleep category. This is called a **completely randomized design**, where each participant is chosen randomly and independently of each other to eliminate bias. For this experiment, we will use a method called the **Analysis of Variance**, also known as ANOVA. There are several types of ANOVA that a statistician can utilize. For our purpose, we will be focusing on **single-factor** (or one-way) ANOVA.

Three basic assumptions are required for ANOVA. First, every group of measurements is obtained from a normal population. Second, each group is randomly selected and independent from every other group. (If a completely randomized design is used, then this assumption is already satisfied.) Lastly, the variables from each group come from distributions with approximately equal standard deviations.

We need to define the following terms: *response variable, factor, treatment,* and *experimental unit*. In the example above with the Algebra test scores, we have two variables – test scores and number of hours of sleep. A response variable is the result that your experiment is measuring. A factor is a variable that affects the outcome of the response variable. In the example, since the objective is to determine whether or not the test score means are the same for the different sleep times, then our response (or dependent) variable is the test score, whereas the factor (or independent variable) is the sleep time. Treatments are the levels of the factor, which means that our treatments are the numbers of hours of sleep (2, 4, 6, and 8). Lastly, experimental units are the objects that the experiment uses, in our case, college students.

The hypotheses for one-way ANOVA are:

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

$$H_a : \text{Not all population means are equal}$$

In other words, for the initial hypothesis, you claim that the treatment population means are equal. For the alternative hypothesis, at least two of the population means are not equal.

An ANOVA table is used to efficiently display and store your calculations. Below is a general form of the ANOVA table.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F***	***p*-value**
Treatments	SSTR	$k-1$	$\text{MSTR} = \frac{\text{SSTR}}{k-1}$	$\frac{\text{MSTR}}{\text{MSE}}$	
Error	SSE	$n_T - k$	$\text{MSE} = \frac{\text{SSE}}{n_T - k}$		
Total	SST	$n_T - 1$			

In this table, k is the number of treatment groups (number of levels of the factor), n_T is the total sample size, and F is the test statistic. The Sum of Squares for Treatments and the Sum of Squares for Error are calculated as follows:

$$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2$$

$$\text{SSE} = \sum_{j=1}^{k} \left(n_j - 1\right) s_j^2$$

$$\text{SST} = \text{SSTR} + \text{SSE}$$

Here, n_j is the sample size of the j$^{\text{th}}$ treatment, $\bar{x}_j$ is the sample mean of the j$^{\text{th}}$ treatment, s_j^2 is the sample variance of the j$^{\text{th}}$ treatment, and $\bar{\bar{x}}$ is the sample mean of all of the treatments combined.

Note, in the ANOVA table above, that there is a Degrees of Freedom column. In the F distribution, two degrees of freedom are required. df_1 is the degrees of freedom of the numerator of the F statistic (MSTR) and df_2 is the degrees of freedom of the denominator of the F statistic (MSE). When you are determining the rejection region for the hypothesis test, the degrees of freedom that you will use in the F table are the first two values in the Degrees of Freedom column. In other words, $df_1 = k - 1$ and $df_2 = n_T - k$.

Interpretation:

The "treatment variation" is the variance between groups – how much group means vary from the overall mean.

The "error variation" is the variance within groups – how much the groups vary about their own means. This is not influenced by the differences between group means.

The F statistic $= \dfrac{\text{between groups variance}}{\text{within groups variance}}$.

If there are no treatment effects (i.e., the null hypothesis is true), this ratio will be close to 1. The higher the F value, the more likely that the difference between groups is real (i.e., the alternative hypothesis is true).

Example 1:

	Treatment		
Observation	**A**	**B**	**C**
1	162	142	126
2	142	156	122
3	165	124	138
4	145	142	140
5	148	136	150
6	174	152	128

Given the data in the table, compute the following:

a. SSTR
b. MSTR
c. SSE
d. MSE
e. Set up the ANOVA table. At the $\alpha = .05$ significance level, test whether the three means are equal.

Solution:

For the three treatments, we get the following means and variances:

	Treatment		
	A	**B**	**C**
Mean	156	142	134
Variance	164.4	131.2	110.4

The total sample mean, $\bar{\bar{x}}$, is 144. Since we have three treatments with six observations each, $k = 3$ and $n_T = 18$. Degrees of freedom are $df_1 = 3 - 1 = 2$ and $df_2 = 18 - 3 = 15$.

a. Using the formula for SSTR:
$$\text{SSTR} = 6(156-144)^2 + 6(142-144)^2 + 6(134-144)^2 = 1488$$

b. $$\text{MSTR} = \frac{\text{SSTR}}{k-1} = \frac{1488}{3-1} = 744$$

c. Using the formula for SSE:
$$\text{SSE} = (6-1)\bullet 164.4 + (6-1)\bullet 131.2 + (6-1)\bullet 110.4 = 2030$$

d. $$\text{MSE} = \frac{\text{SSE}}{n_T - k} = \frac{2030}{18-3} \approx 135.3$$

e. $$F = \frac{\text{MSTR}}{\text{MSE}} \approx \frac{744}{135.3} \approx 5.50.$$

Now we can fill in the table. To find the rejection region for the test, look in the F distribution table with $\alpha = .05$, $df_1 = 2$, and $df_2 = 15$; this gives you a rejection region of $F > 3.68$. Therefore, we reject H_0. In other words, at least two of the population means are different.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F***	***p*-value**
Treatments	1488	2	744	5.50	0.0162*
Error	2030	15	135.3		
Total	3518	17			

Example 2:

Complete the following ANOVA table and test using $\alpha = .05$ (refer to the table on page 2). For each of the five treatments, seven experimental units were sampled.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F***	***p*-value**
Treatments	300				
Error					
Total	460				

Since there were five treatments and seven units, then $k = 5$ and $n_j = 7$, respectively. This means that the total sample size is $n_T = 5 \cdot 7 = 35$. So, we now have the following table:

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F***	***p*-value**
Treatments	300	4			
Error		30			
Total	460	34			

Finishing the computations, we get the final table.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F***	***p*-value**
Treatments	300	4	75	14.1	0.0000014*
Error	160	30	5.33		
Total	460	34			

For $\alpha = .05$, $df_1 = 4$, $df_2 = 30$, the rejection region is $F > 2.69$. Therefore, we reject the null hypothesis.

* For the *p*-value, use the following command on the TI-83/84:
2^{nd} → DISTR → Fcdf(F value, 1EE99, df_1, df_2)