

Chapter 11

11. Analysis of Variance (ANOVA)

11.1. Introduction

The F test used to compare three or more means. This technique is called analysis of variance, or ANOVA. It is used to test claims involving three or more means. (Note: The F test can also be used to test the equality of two means. But since it is equivalent to the t test in this case, the t test is usually used instead of the F test when there are only two means.) The researcher will use the ANOVA technique for this test. The z and t tests should not be used when three or more means are compared.

The analysis of variance that is used to compare three or more means is called a one way analysis of variance since it contains only one variable.

11.2 One-Way Analysis of Variance

When an F test is used to test a hypothesis concerning the means of three or more populations, the technique is called analysis of variance (commonly abbreviated as ANOVA).

At first glance, you might think that to compare the means of three or more samples, you can use the t test, comparing two means at a time. But there are several reasons why the t test should not be done.

First, when you are comparing two means at a time, the rest of the means under study are ignored. With the F test, all the means are compared simultaneously. Second, when you are comparing two means at a time and making all pair wise comparisons, the probability of rejecting the null hypothesis when it is true is increased, since the more t tests that are conducted, the greater is the likelihood of getting significant differences by chance alone. Third, the more means there are to compare, the more t tests are needed.

Assumptions for the F Test for comparing three or more means

1. The populations from which the samples were obtained must be normally or approximately normally distributed.
2. The samples must be independent of one another.
3. The variances of the populations must be equal.

Even though you are comparing three or more means in this use of the F test, variances are used in the test instead of means. With the F test, two different estimates of the population variance are made. The first estimate is called the between-group variance, and it involves finding the variance of the means. The second estimate, the within-group variance, is made by computing the variance using all the data and is not affected by differences in the means. If there is no difference in the means, the between-group variance estimate will be approximately equal to the within-group variance estimate, and the F test value will be approximately equal to 1. The null hypothesis will not be rejected. However, when the means differ significantly, the between-group variance will be much larger than the within-group variance; the F test value will be significantly greater than 1; and the null hypothesis will be rejected. Since variances are compared, this procedure is called analysis of variance (ANOVA). For a test of the difference among three or more means, the following hypotheses should be used:

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

H_1 : At least one mean is different from the others.

The degrees of freedom for this F test are d.f.N. = $k-1$, where k is the number of groups, and d.f.D. = $N-k$, where N is the sum of the sample sizes of the groups $N = n_1 + n_2 + \dots + n_k$. The sample sizes need not be equal. The F test to compare means is always right-tailed.

Example 1: A researcher wishes to try three different techniques to lower the blood pressure of individuals diagnosed with high blood pressure. The subject are randomly assigned to three groups; the first group takes medication, the second group exercise, and the third group follows a special diet. After four weeks, the reduction in each person's blood pressure is recorded.

At $\alpha = 0.05$, test the claim that there is no difference among the means. The data are shown:

Medication	Exercise	Diet
10	6	5
12	8	9
9	3	12
15	0	8
13	2	4
$\bar{X}_1 = 11.8$	$\bar{X}_2 = 3.8$	$\bar{X}_3 = 7.6$
$s_1^2 = 5.7$	$s_2^2 = 10.2$	$s_3^2 = 10.3$

Solution

Step 1 State the hypotheses and identify the claim.

$$H_0: \mu_1 = \mu_2 = \mu_3 \text{ (claim)}$$

H_1 : At least one mean is different from the others.

Step 2 Find the critical value. Since $k = 3$ and $N = 15$,

$$\text{d.f.N.} = k - 1 = 3 - 1 = 2$$

$$\text{d.f.D.} = N - k = 15 - 3 = 12$$

The critical value ($F_{\alpha=0.05, k-1=2, N-k=12}$) = 3.89 which is a table value.

Step 3 Compute the test value, using the procedure outlined here.

- Find the mean and variance of each sample (these values are shown below the data).
- Find the grand mean. The *grand mean*, denoted by \bar{X}_{GM} , is the mean of all values in the samples.

$$\bar{X}_{GM} = \frac{\sum X}{N} = \frac{10 + 12 + 9 + \dots + 4}{15} = \frac{116}{15} = 7.73$$

When samples are equal in size, find \bar{X}_{GM} by summing the \bar{X} 's and dividing by k , where k = the number of groups.

c. Find the between-group variance, denoted by s_B^2 .

$$\begin{aligned} s_B^2 &= \frac{\sum n_i(\bar{X}_i - \bar{X}_{GM})^2}{k - 1} \\ &= \frac{5(11.8 - 7.73)^2 + 5(3.8 - 7.73)^2 + 5(7.6 - 7.73)^2}{3 - 1} \\ &= \frac{160.13}{2} = 80.07 \end{aligned}$$

Note: This formula finds the variance among the means by using the sample sizes as weights and considers the differences in the means.

d. Find the within-group variance, denoted by s_W^2 .

$$\begin{aligned} s_W^2 &= \frac{\sum (n_i - 1)s_i^2}{\sum (n_i - 1)} \\ &= \frac{(5 - 1)(5.7) + (5 - 1)(10.2) + (5 - 1)(10.3)}{(5 - 1) + (5 - 1) + (5 - 1)} \\ &= \frac{104.80}{12} = 8.73 \end{aligned}$$

Note: This formula finds an overall variance by calculating a weighted average of the individual variances. It does not involve using differences of the means.

e. Find the F test value.

$$F = \frac{s_B^2}{s_W^2} = \frac{80.07}{8.73} = 9.17$$

Step 4 Make the decision. The decision is to reject the null hypothesis, since $9.17 > 3.89$.

Step 5 Summarize the results. There is enough evidence to reject the claim and conclude that at least one mean is different from the others.

The numerator of the fraction obtained in step 3, part c, of the computational procedure is called the sum of squares between groups, denoted by SS_B . The numerator of the fraction obtained in step 3, part d, of the computational procedure is called the sum of squares within groups, denoted by SS_W . This statistic is also called the sum of squares for the error.

SS_B is divided by d.f.N. to obtain the between-group variance. SS_W is divided by $N - k$ to obtain the within-group or error variance. These two variances are sometimes called mean squares,

denoted by MS_B and MS_W . These terms are used to summarize the analysis of variance and are placed in a summary table, as shown in the following Table.

Source	Sum of squares	df	Mean Square	F
Between	SS_B	k-1	MS_B	MS_B/MS_W
Within error	SS_W	N-k	MS_W	
Total	SS_T			

In the table,

SS_B = sum of squares between groups

SS_W = sum of squares within groups

k = number of groups

$N = n_1 + n_2 + \dots + n_k$ = sum of sample sizes for groups

$$MS_B = \frac{SS_B}{k - 1}$$

$$MS_W = \frac{SS_W}{N - k}$$

$$F = \frac{MS_B}{MS_W}$$

The totals are obtained by adding the corresponding columns. For Example 1, the ANOVA summary table is shown in Table 2.

Source	Sum of squares	df	Mean Square	F
Between	160.13	3-1=2	80.07	$80.07/8.73 = 9.17$
Within error	104.80	14-2=12	8.73	
Total	264.93	14		

11.3 ANOVA VS Regression

Both the Regression and ANOVA are the statistical models which are used in order to predict the continuous outcome but in case of the regression, continuous outcome is predicted on basis of

the one or more than one continuous predictor variables whereas in case of ANOVA continuous outcome is predicted on basis of the one or more than one categorical predictor variables.

Regression is a statistical method to establish the relationship between sets of variables in order to make predictions of the dependent variable with the help of independent variables, ANOVA, on the other hand, is a statistical tool applied on unrelated groups to find out whether they have a common mean.

- Regression is applied to variables that are mostly fixed or independent in nature and ANOVA is applied to random variables.
- Regression is mainly used in two forms they are linear regression and multiple regression, though other forms of regression are also present in theory those types are most widely used in practice, on the other hand, there are three popular types of ANOVA they are a random effect, fixed effect, and mixed effect.
- Regression is mainly used in order to make estimates or predictions for the dependent variable with the help of single or multiple independent variables and ANOVA is used to find a common mean between variables of different groups.
- In the case of regression, the number of the error term is one but in the case of ANOVA, the number of the error term is more than one.

Comparative Table

Basis	Regression	ANOVA
Definition	Regression is a very effective statistical method to establish the relationship between sets of variables.	ANOVA is the short form of analysis of variance. It is applied to unrelated groups to find out whether they have a common mean
Nature of Variable	Regression is applied on independent variables or fixed variables.	ANOVA is applied to variables which are random in nature
Types	Regression is mainly used in two forms they are linear regression and multiple	The three popular types of ANOVA are a random effect, fixed effect, and

	regression, the latter is when the number of independent variables is more than one.	mixed effect.
Examples	A paint company uses solvent & monomers as its raw material which is a derivative of crude; we can run a <u>regression analysis</u> between the price of that raw material and the price of Brent crude prices.	If two separate research teams are researching different products not related to each other. ANOVA will help to find which one is providing better results.
Variables Used	Regression is applied to two sets of variables, one of them is the dependent variable and the other one is the independent variable. The number of independent variables in regression can be one or more than one.	ANOVA is applied to variables from different which not necessarily related to each other.
Use of the Test	Regression is mainly used by the practitioners or industry experts in order to make estimates or predictions for the dependent variable.	ANOVA is used to find a common mean between variables of different groups.
Errors	The predictions made by the regression analysis are not always desirable that's because of the error term in a regression, this error term is also known as residual. In the case of regression, the number of the error term is one.	The number of errors in case ANOVA, unlike regression, is more than one.

Both regressions and ANOVA are powerful statistical tools that are applied to multiple variables. Regression is used in order to make predictions of the dependent variable with the help of

independent variables that have some relations. It is helpful to validate a hypothesis of whether the hypothesis made is correct or not.

Regression is used on variables that are fixed or independent in nature and can be done with the use of a single independent variable or multiple independent variables. ANOVA is used to find a common between variables of different groups that are not related to each other. It is not used to make a prediction or estimate but to understand the relations between the set of variables.

11.4 Multiple Comparisons

11.4.1 Introduction:

A significant F-test tells us that at least one of the underlying population means are different, but it does not tell us which ones differ from the others. We need extra tests to compare all the means, which we call Multiple Comparisons. We look at the difference between every pair of group population means, as well as the confidence interval for each difference.

If the overall F statistic from the ANOVA is statistically significant, multiple comparisons procedures can be used in an attempt to discover the source of the significant differences among the group means. There are many different multiple comparison procedures.

11.4.2. Fisher's Least significant difference (LSD) and Scheffe's S Method

Fisher's Least significant difference (LSD)

When an ANOVA gives a significant result, this indicates that at least one group differs from the other groups. Yet, the omnibus test does not indicate which group differs. In order to analyze the pattern of difference between means, the ANOVA is often followed by specific comparisons, and the most commonly used involves comparing two means ("pair wise comparison"). This technique can be used only if the ANOVA F omnibus is significant.

The main idea of the LSD is to compute the smallest significant difference (i.e., the LSD) between two means as if these means had been the only means to be compared (i.e., with a t - test) and to declare significant any difference larger than the LSD.

Procedure

1. Prefer an analysis of variance to test $H_0: \mu_1 = \mu_2 = \dots = \mu_t$ Vs H_1 : at least one of the means differs from the rest.
2. If there is insufficient evidence to reject H_0 using $F = BMS/WMS$, proceed no further.
3. If H_0 is rejected, define the **least significant difference (LSD)** to be the observed difference between two sample means necessary to declare the corresponding population means different.
4. For a specified value of α , the least significant difference for comparing μ_i to μ_j is

$$LSD_{ij} = t_{\alpha/2} \sqrt{WMS \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

, where n_i and n_j are the respective sample sizes from population i and j and t is the critical value for $\alpha = \alpha/2$ and df denoting the degrees of freedom for WMS .

Note that for $n_1 = n_2 = \dots n_t = n$

$$LSD_{ij} = t_{\alpha/2} \sqrt{2WMS/n}$$

5. Then compare all pairs of sample means. If $|\bar{x}_i - \bar{x}_j| \geq LSD_{ij}$, declare the corresponding population means μ_i and μ_j different.
6. For each pair wise comparison of population means, the probability of a type I error is fixed at a specified value of α .

Determining significance

The difference between two mean values is compared to the LSD value. If the difference is greater than the LSD value, then the means are significantly different.

Example: We want to compare the mean leaf width of a new variety (A) with two comparators (B & C). The data is presented as follows:

Variety	Variety A	17	15	10	14
	Variety B	34	26	23	22
	Variety C	23	21	8	16

Solution: We can solve this problem by following the five steps listed for the LSD procedure.

$H_0: \mu_1 = \mu_2 = \mu_3$ Vs H_1 : at least one of the means differs from the rest.

First, calculate the variety totals and means. Then, calculate the replicate totals.

Variety	Sample Size	Sample Mean	Sample variance
A	$n_1 = 4$	14	$s_1^2 = 8.6667$
B	$n_1 = 4$	26.25	$s_2^2 = 29.5833$
C	$n_1 = 4$	17	$s_3^2 = 44.6667$

The grand mean is:

$$\bar{X}_{..} = \frac{\sum_{i=1}^3 \sum_{j=1}^4 X_{ij}}{N} = \frac{\sum_{i=1}^3 n_i \bar{x}_i}{12} = \frac{4(14) + 4(26.25) + 4(17)}{12} = 19.08$$

$$\begin{aligned} BSS &= \sum_{i=1}^3 \sum_{j=1}^4 (\bar{X}_i - \bar{X}_{..})^2 = \sum_{i=1}^3 n_i (\bar{X}_i - \bar{X}_{..})^2 = 4(14 - 19.08)^2 + 4(26.25 - \\ &19.08)^2 + 4(17 - 19.08)^2 \\ &= 326.1668 \end{aligned}$$

$$WSS = \sum_{i=1}^3 (n_i - 1) s_i^2 = 3 \times 8.6667 + 3 \times 29.5833 + 3 \times 44.6667 = 248.7501$$

$$TSS = BSS + WSS = 326.1668 + 248.7501 = 574.9169$$

ANOVA table:

Source of variation	Df.	SS	MSS	F- test
Between treatments	2	326.1668	163.0834	5.9005
Within treatments (error)	9	248.7501	27.6389	
Total	11	574.9169		

Critical value: $F_{0.05, 2, 9} = 4.26$

Decision: reject the null hypothesis, since the calculated value is greater than the tabulated value.

Conclusion: The calculated F value 5.9005 is more than the table F value 4.26. This reveals that there are significant differences among the three varieties i.e. at least one variety is different from any of others. The next step is to calculate LSD

The least significant difference for comparing two means based on samples of size 4 is then

$$LSD = t_{\alpha/2} \sqrt{2MS_{Error/n}} = 2.262 \sqrt{2(27.6389)/4} = 8.4089$$

Note that the appropriate t value (2.262) was obtained from table value with $\alpha = \alpha/2 = 0.025$ and df 9.

Step5. When we have equal sample size, it is convenient to use the following procedures rather than make all pair wise comparison among the sample means because the same LSD is to be used for all comparisons.

Determining if the two varietal means are significantly different:

The difference between two mean values is compared to the LSD value. If the difference is greater than the LSD value, then the means are significantly different. The variety means for the above example are A (candidate) = 14, B (comparator 1) = 26.25, C (comparator 2) = 17.

The absolute difference between A and B is 12.25 which is greater than 8.4089: therefore A and B are significantly different at $P \leq 0.05$. This confirms the results of F test.

The difference between A and C is 3 which is less than 8.4089: therefore A and C are not significantly different at $P \leq 0.05$.

The difference between B and C is 9.25 which is greater than 8.4089: therefore A and C are significantly different at $P \leq 0.05$.

Scheffe's S Method

To conduct the **Scheffé test**, you must compare the means two at a time, using all possible combinations of means. For example, if there are three means, the following comparisons must be done:

$$\bar{X}_1 \text{ versus } \bar{X}_2 \quad \bar{X}_1 \text{ versus } \bar{X}_3 \quad \bar{X}_2 \text{ versus } \bar{X}_3$$

$$F_S = \frac{(\bar{X}_i - \bar{X}_j)^2}{S_W^2 \left[\frac{1}{n_i} + \frac{1}{n_j} \right]}$$

Where \bar{X}_i and \bar{X}_j are the means of the samples being compared, n_i and n_j are the respective sample sizes, and S_W^2 is the within-group variance.

To find the critical value F' for the Scheffé test, multiply the critical value for the F test by $k-1$:

$$F' = (k-1)(C.V.)$$

There is a significant difference between the two means being compared when F_S is greater than F' .

Using the Scheffé test, test each pair of means in Example 2 to see whether a specific difference exists, at $\alpha = 0.05$.

Solution:

i) for \bar{X}_1 versus \bar{X}_2 ,

$$F_S = \frac{(\bar{X}_1 - \bar{X}_2)^2}{S_W^2 \left[\frac{1}{n_1} + \frac{1}{n_2} \right]} = \frac{(11.8 - 3.8)^2}{8.73 \left[\frac{1}{5} + \frac{1}{5} \right]} = 18.33$$

ii) for \bar{X}_2 versus \bar{X}_3 ,

$$F_S = \frac{(\bar{X}_2 - \bar{X}_3)^2}{S_W^2 \left[\frac{1}{n_2} + \frac{1}{n_3} \right]} = \frac{(3.8 - 7.6)^2}{8.73 \left[\frac{1}{5} + \frac{1}{5} \right]} = 4.14$$

iii) for \bar{X}_1 versus \bar{X}_3 ,

$$F_s = \frac{(\bar{X}_1 - \bar{X}_3)^2}{S_w^2 \left[\frac{1}{n_1} + \frac{1}{n_3} \right]} = \frac{(11.8 - 7.6)^2}{8.73 \left[\frac{1}{5} + \frac{1}{5} \right]} = 5.05$$

The critical value for the analysis of variance; by using table value with $\alpha = 0.05$, d.f.N = $k-1=2$, and d.f.D = $N-k = 12$, is 3.89. The critical value for F' at $\alpha = 0.05$, with d.f.N = 2 and d.f.D. = 12 is:

$$F' = (k-1)(C.V) = (3-1)(3.89) = 7.78$$

Since only the F test value for part i (\bar{X}_1 versus \bar{X}_2) is greater than the critical value, 7.78, the only significant difference is between \bar{X}_1 and \bar{X}_2 , that is, between medication and exercise.