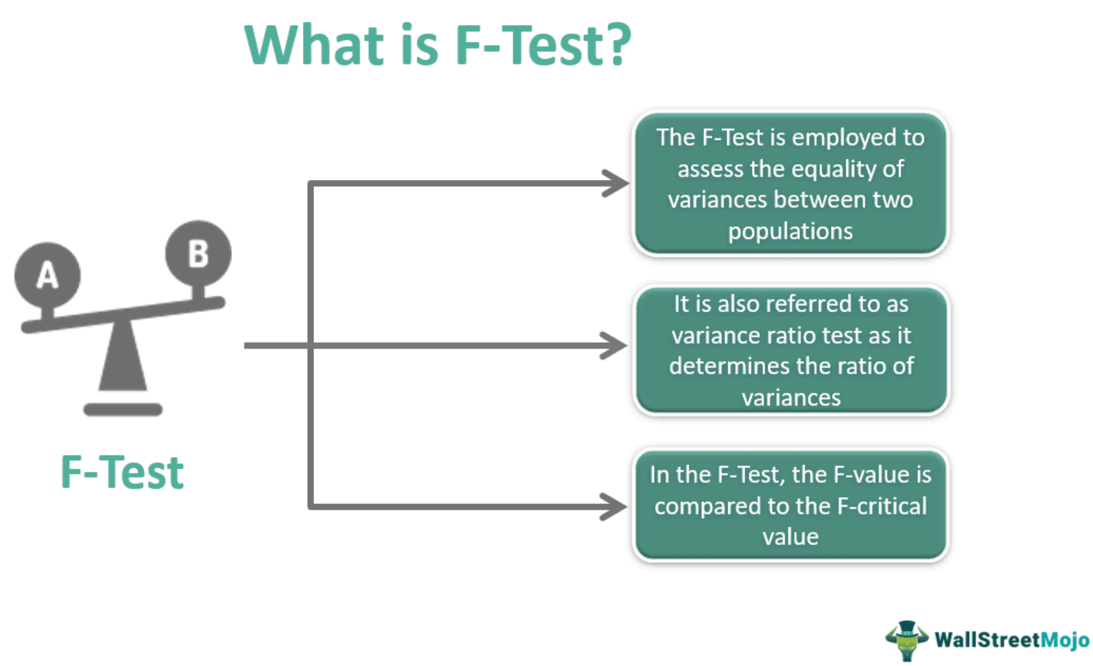
UNIT-3

Analysis of Variance

F-Test

F-Test in statistics is a hypothesis-testing procedure that considers two variances from two samples. The F-Test is used when the difference between two variances needs to be significantly assessed, i.e., when determining whether or not two samples can be taken as representative of the normal population with the same variance.



The F-Test helps to determine the overall significance of the regression. It is useful in various situations, such as when a quality controller wants to determine whether the product's quality is deteriorating over time. In addition, it might be useful for an economist to determine whether income variability varies between two populations.

Points:

- The F-test is a statistical test that evaluates if the variances of the two normal populations are equal.
- One can deem the variance ratio of the test insignificant if F OR = F0.5, and one can assume that the values will be from the same group or groups with similar variances.
- The null hypothesis is rejected, and the variance ratio is considered significant if F> OR = F0.5.
- The F-test vs. t-test: The t-test and the F-test are two separate tests. The T-test compares two populations' means, whereas the other compares two populations' variances.

The following conditions are critical for using the F-test to compare the variances of two populations:

1. **Normality:** the populations must have a **normal distribution**.
2. **Independent and random selection of sample items:** the selection of the samples' components should be independent and random.
3. **More than unity:** The variance ratio must be one or larger than one; it cannot be less than one. When dividing variance estimates, smaller estimates divide the larger estimates of variances.
4. The additive property states that the total of different variance components will equal the total variance, i.e., the total variance is the sum of the variance between samples and the variance within samples.

Formula

1. **Sample variances:** The formula for calculating sample variances is as follows (an online F-test calculator can make it easier):

$$S_1{}^2 = \frac{\sum (x_1 - \overline{x_1})^2}{n_1 - 1} \qquad S_2{}^2 = \frac{\sum (x_2 - \overline{x_2})^2}{n_2 - 1}$$

2. **Null hypothesis:** After the formation of the test, the **null hypothesis** are either

a) Two samples were from the same group or

(b) The population's variances concerning both samples are equal.

3. To compute the variance ratio, use the formula F = larger estimate divided by a smaller estimate of variance. Regardless of whether S12 or S22, the numerator will always be the larger value.

4. When calculating **degrees of freedom**, the larger the sample's variance is V1; the smaller variance is V2.

5. **Table value of F:** the critical value of F is available from the "F-Table" (F-test table) at the determined significance level.

6. **Analysis:** This involves the comparison of the computed value and the tabulated value. For various levels of significance, there are several F Tables (F-test tables).

(a) The variance ratio is insignificant if F < OR = F0.5. We can assume that the values are from the same group or groups with similar variances.

(b) The null hypothesis is rejected, and the variance ratio is considered significant if F> OR = F0.5.

Formula

1. **Sample variances:** The formula for calculating sample variances is as follows (an online F-test calculator can make it easier):

$$S_1^2 = \frac{\sum (x_1-\overline{x_1})^2}{n_1-1} \qquad S_2^2 = \frac{\sum (x_2-\overline{x_2})^2}{n_2-1}$$

2. **Null hypothesis:** After the formation of the test, the **null hypothesis** are either

a) Two samples were from the same group or

(b) The population's variances concerning both samples are equal.

3. To compute the variance ratio, use the formula F = larger estimate divided by a smaller estimate of variance. Regardless of whether S12 or S22, the numerator will always be the larger value.

4. When calculating **degrees of freedom**, the larger the sample's variance is V1; the smaller variance is V2.

5. **Table value of F:** the critical value of F is available from the "F-Table" (F-test table) at the determined significance level.

6. **Analysis:** This involves the comparison of the computed value and the tabulated value. For various levels of significance, there are several F Tables (F-test tables).

(a) The variance ratio is insignificant if F < OR = F0.5. We can assume that the values are from the same group or groups with similar variances.

(b) The null hypothesis is rejected, and the variance ratio is considered significant if F> OR = F0.5.

Analysis of variance(ANOVA)

An analysis of variance (ANOVA) tests whether statistically significant differences exist between more than two samples. For this purpose, the means and variances of the respective groups are compared with each other. In contrast to the t-test, which tests whether there is a difference between two samples, the ANOVA tests whether there is a difference between more than two groups.

There are different types of analysis of variance, being the one-way and two-way analyses of variance the most common ones, each of which can be calculated either with or without repeated measurements.

Assumptions for ANOVA

1. The dependent variable is approximately normally distributed within each group. This assumption is more critical for smaller sample sizes.
2. The samples are selected at random and should be independent of one another.
3. All groups have equal standard deviations.
4. Each data point should belong to one and only one group. There should be no overlap or sharing of data points between groups.

In the case of two-way ANOVA, there are additional assumptions related to the interaction between the independent variables.

- The effect of one independent variable on the dependent variable should be consistent across all levels of the other independent variable.
- Combined effect of two independent variables is equal to the sum of their individual effects.

Types of ANOVA

There are two main types of ANOVA:

1. **One-way ANOVA**: This is the most basic form of ANOVA and is used when there is only one independent variable with more than two levels or groups. It assesses whether there are any statistically significant differences among the means of the groups.
2. **Two-way ANOVA**: Extending the one-way ANOVA, two-way ANOVA involves two independent variables. It allows for the examination of the main effects of each variable as well as the interaction between them. The interaction effect explores whether the effect of one variable on the dependent variable is different depending on the level of the other variable.

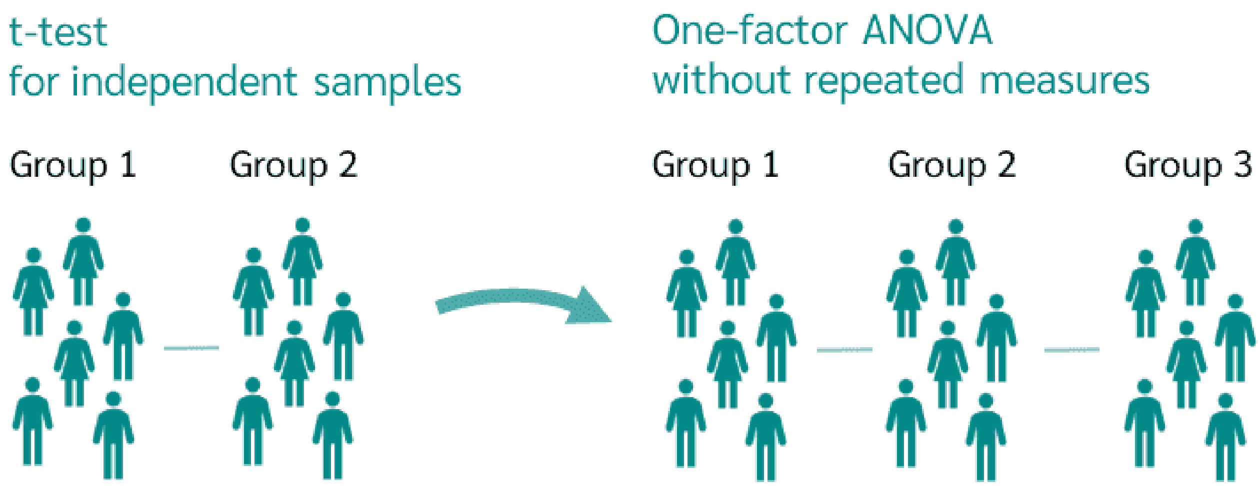
Why not calculate multiple t-tests?

ANOVA is used when there are more than two groups. Of course, it would also be a possibility to calculate a t-test for each combination of the groups. The problem here, however, is that every hypothesis test has some degree of error. This probability of error is usually set at 5%, so that, from a purely statistical point of view, every 20^{th} test gives a wrong result

If, for example, 20 groups are compared in which there is actually no difference, one of the tests will show a significant difference purely due to the sampling.

One-factorial Analysis of Variance (One-way ANOVA)

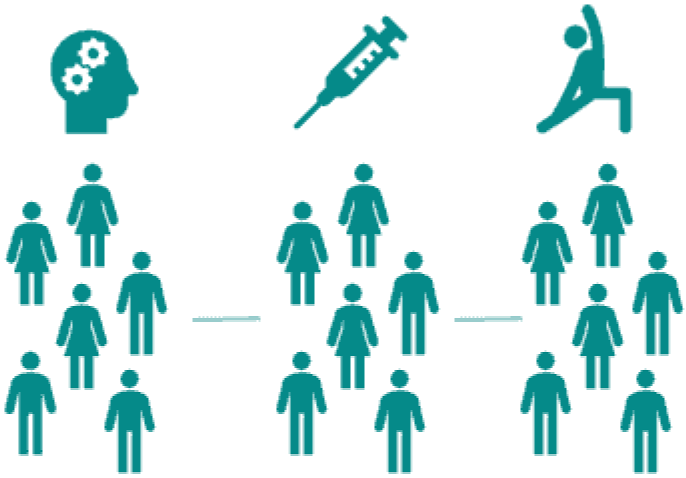
The one factorial analysis of variance tests whether there is a difference between the means of more than 2 groups. Thus, one-way ANOVA is the extension of the independent t-test to more than two groups or samples.



One Factor ANOVA Example

A classic use case for analysis of variance is in therapy research. For example, you might be interested in whether different therapies result in different therapeutic successes after a herniated disc. For this we could test three different therapies.

On the one hand you could just discuss with the patient which movements are good and which are bad for the disc, then you could treat one group with medication and with the last group you could do stretching and strength training.



At the end of the therapy, we could then measure the success and use an analysis of variance to calculate whether there is a significant difference between the three types of therapy. Of course, the assumptions have to be fulfilled in order to calculate an ANOVA, more about this later.

Two way ANOVA:

A two-way ANOVA is used to estimate how the mean of a quantitative variable changes according to the levels of two categorical variables. Use a two-way ANOVA when you want to know how two independent variables, in combination, affect a dependent variable.

Example

We are researching which type of fertilizer and planting density produces the greatest crop yield in a field experiment. We assign different plots in a field to a combination of fertilizer type (1, 2, or 3) and planting density (1=low density, 2=high density), and measure the final crop yield in bushels per acre at harvest time.

We can use a two-way ANOVA to find out if fertilizer type and planting density have an effect on average crop yield.

When to use a two-way ANOVA

We can use a two-way ANOVA when we have collected data on a quantitative dependent variable at multiple levels of two categorical independent variables.

A **quantitative variable** represents amounts or counts of things. It can be divided to find a group mean.

A **categorical variable** represents types or categories of things. A level is an individual category within the categorical variable.

Fertilizer types 1, 2, and 3 are levels within the categorical variable **fertilizer type**. Planting densities 1 and 2 are levels within the categorical variable **planting density**.

Assumptions of the two-way ANOVA

To use a two-way ANOVA our data should meet certain assumptions. Two-way ANOVA makes all of the normal assumptions of a parametric test of difference:

1. **Homogeneity of variance (a.k.a.** homoscedasticity**)**

The variation around the mean for each group being compared should be similar among all groups.

Independence of observations

our independent variables should not be dependent on one another (i.e. one should not cause the other). This is impossible to test with categorical variables – it can only be ensured by good experimental design.

In addition, our dependent variable should represent unique observations – that is, our observations should not be grouped within locations or individuals.

3. **Normally-distributed dependent variable**

The values of the dependent variable should follow a bell curve (they should be normally distributed).

In the crop-yield example, the response variable is normally distributed, and we can check for homoscedasticity after running the model. The experimental treatments were set up within blocks in the field, with four blocks each containing every possible combination of fertilizer type and planting density, so we should include this as a blocking variable in the model.

Summary: differences between one-way and two-way ANOVA

The key differences between one-way and two-way ANOVA are summarized clearly below.

1. A one-way ANOVA is primarily designed to enable the equality testing between three or more means. A two-way ANOVA is designed to assess the interrelationship of two independent variables on a dependent variable.

2. A one-way ANOVA only involves one factor or independent variable, whereas there are two independent variables in a two-way ANOVA.

3. In a one-way ANOVA, the one factor or independent variable analyzed has three or more categorical groups. A two-way ANOVA instead compares multiple groups of two factors.

4. One-way ANOVA need to satisfy only two principles of design of experiments, i.e. replication and randomization. As opposed to two-way ANOVA, which meets all three principles of design of experiments which are replication, randomization and local control.

One-way vs two-way ANOVA differences chart

	One-Way ANOVA	Two-Way ANOVA
Definition	A test that allows one to make comparisons between the means of three or more groups of data.	A test that allows one to make comparisons between the means of three or more groups of data, where two independent variables are considered.
Number of Independent Variables	One.	Two.
What is Being Compared?	The means of three or more groups of an independent variable on a dependent variable.	The effect of multiple groups of two independent variables on a dependent variable and on each other.
Number of Groups of Samples	Three or more.	Each variable should have multiple samples.