What is Analysis of Variance (ANOVA)?

Analysis of Variance (**ANOVA**) is a statistical formula used to compare variances across the means (or average) of different groups. A range of scenarios use it to determine if there is any difference between the means of different groups.

For example, to study the effectiveness of different diabetes medications, scientists design and experiment to explore the relationship between the type of medicine and the resulting blood sugar level. The sample population is a set of people. We divide the sample population into multiple groups, and each group receives a particular medicine for a trial period. At the end of the trial period, blood sugar levels are measured for each of the individual participants. Then for each group, the mean blood sugar level is calculated. ANOVA helps to compare these group means to find out if they are statistically different or if they are similar.

The outcome of ANOVA is the 'F statistic'. This ratio shows the difference between the within group variance and the between group variance, which ultimately produces a figure which allows a conclusion that the null hypothesis is supported or rejected. If there is a significant difference between the groups, the null hypothesis is not supported, and the F-ratio will be larger.

ANOVA Terminology

Dependent variable: This is the item being measured that is theorized to be affected by the independent variables.

Independent variable/s: These are the items being measured that may have an effect on the dependent variable.

A null hypothesis (H0): This is when there is no difference between the groups or means. Depending on the result of the ANOVA test, the null hypothesis will either be accepted or rejected.

An alternative hypothesis (H1): When it is theorized that there is a difference between groups and means.

Factors and levels: In ANOVA terminology, an independent variable is called a factor which affects the dependent variable. Level denotes the different values of the independent variable that are used in an experiment.

Fixed-factor model: Some experiments use only a discrete set of levels for factors. For example, a fixed-factor test would be testing three different dosages of a drug and not looking at any other dosages.

Random-factor model: This model draws a random value of level from all the possible values of the independent variable.

What is the Difference Between One Factor and Two Factor ANOVA?

There are two types of ANOVA.

One-Way ANOVA

The one-way analysis of variance is also known as single-factor ANOVA or simple ANOVA. As the name suggests, the one-way ANOVA is suitable for experiments with only one independent variable (factor) with two or more levels. For instance a dependent variable may be what month of the year there are more flowers in the garden. There will be twelve levels. A one-way ANOVA assumes:

- Independence: The value of the dependent variable for one observation is independent of the value of any other observations.
- Normalcy: The value of the dependent variable is normally distributed
- Variance: The variance is comparable in different experiment groups.
- Continuous: The dependent variable (number of flowers) is continuous and can be measured on a scale which can be subdivided.

Full Factorial ANOVA (also called two-way ANOVA)

Full Factorial ANOVA is used when there are two or more independent variables. Each of these factors can have multiple levels. Full-factorial ANOVA can only be used in the case of a full factorial experiment, where there is use of every possible permutation of factors and their levels. This might be the month of the year when there are more flowers in the garden, and then the number of sunshine hours. This two-way ANOVA not only measures the independent vs the independent variable, but if the two factors affect each other. A two-way ANOVA assumes:

- Continuous: The same as a one-way ANOVA, the dependent variable should be continuous.
- Independence: Each sample is independent of other samples, with no crossover.
- Variance: The variance in data across the different groups is the same.
- Normalcy: The samples are representative of a normal population.
- Categories: The independent variables should be in separate categories or groups.

Why Does ANOVA work?

Some people question the need for ANOVA; after all, mean values can be assessed just by looking at them. But ANOVA does more than only comparing means.

Even though the mean values of various groups appear to be different, this could be due to a sampling error rather than the effect of the independent variable on the dependent variable. If it is due to sampling error, the difference between the group means is meaningless. ANOVA helps to find out if the difference in the mean values is statistically significant.

ANOVA also indirectly reveals if an independent variable is influencing the dependent variable. For example, in the above blood sugar level experiment, suppose ANOVA finds that group means are not statistically significant, and the difference between group means is only due to sampling error. This result infers that the type of medication (independent variable) is not a significant factor that influences the blood sugar level.

Limitations of ANOVA

ANOVA can only tell if there is a significant difference between the means of at least two groups, but it can't explain which pair differs in their means. If there is a requirement for granular data, deploying further follow up statistical processes will assist in finding out which groups differ in mean value. Typically, ANOVA is used in combination with other statistical methods.

ANOVA also makes assumptions that the dataset is uniformly distributed, as it compares means only. If the data is not distributed across a normal curve and there are outliers, then ANOVA is not the right process to interpret the data.

Similarly, ANOVA assumes the standard deviations are the same or similar across groups. If there is a big difference in standard deviations, the conclusion of the test may be inaccurate.

How is ANOVA Used in Data Science?

One of the biggest challenges in machine learning is the selection of the most reliable and useful features that are used in order to train a model. ANOVA helps in selecting the best features to train a model. ANOVA minimizes the number of input variables to reduce the complexity of the model. ANOVA helps to determine if an independent variable is influencing a target variable.

An example of ANOVA use in data science is in email spam detection. Because of the massive number of emails and email features, it has become very difficult and resource-intensive to identify and reject all spam emails. ANOVA and f-tests are deployed to identify features that were important to correctly identify which emails were spam and which were not.

Questions That ANOVA Helps to Answer

Even though ANOVA involves complex statistical steps, it is a beneficial technique for businesses via use of AI. Organizations use ANOVA to make decisions about which alternative to choose among many possible options. For example, ANOVA can help to:

- Compare the yield of two different wheat varieties under three different fertilizer brands.
- Compare the effectiveness of various social media advertisements on the sales of a particular product.
- Compare the effectiveness of different lubricants in different types of vehicles.

What is ANCOVA?

ANCOVA is a blend of analysis of variance (ANOVA) and regression. It is similar to factorial ANOVA, in that it can tell you what additional information you can get by considering one independent variable (factor) at a time, without the influence of the others. It can be used as:

- An extension of multiple regression to compare multiple regression lines,
- An extension of analysis of variance.

Although ANCOVA is usually used when there are differences between your baseline groups (Senn, 1994; Overall, 1993), it can also be used in pretest/posttest analysis when regression to the mean affects your posttest measurement (Bonate, 2000). The technique is also common in non-experimental research (e.g. surveys) and for quasi-experiments (when study participants can't be assigned randomly). However, this particular application of ANCOVA is not always recommended (Vogt, 1999).

Extension of Multiple Regression

When used as an extension of multiple regression, ANCOVA can test all of the regression lines to see which have different Y intercepts as long as the slopes for all lines are equal.

Like regression analysis, ANCOVA enables you to look at how an independent variable acts on a dependent variable. ANCOVA **removes any effect of covariates,** which are variables you don't want to study. For example, you might want to study how different levels of teaching skills affect student performance in math; It may not be possible to randomly assign students to classrooms. You'll need to account for systematic differences between the students in different classes (e.g. different initial levels of math skills between gifted and mainstream students).

Example

You might want to find out if a new drug works for depression. The study has three treatment groups and one control group. A regular ANOVA can tell you if the treatment works. ANCOVA can control for other factors that might influence the outcome. For example: family life, job status, or drug use.

Extension of ANOVA

As an extension of ANOVA, ANCOVA can be used in two ways (Leech et. al, 2005):

1. To control for covariates (typically continuous or variables on a particular scale) that aren't the main focus of your study.
2. To study combinations of categorical and continuous variables, or variables on a scale as predictors. In this case, the covariate is a variable of interest (as opposed to one you want to control for).

Within-Group Variance

ANCOVA can explain within-group variance. **It takes the unexplained variances from the ANOVA test and tries to explain them** with confounding variables (or other covariates). You can use multiple possible covariates. However, more you enter, the fewer degrees of freedom you'll have. Entering a weak covariate *isn't* a good idea as it will reduce the statistical power. The lower the power, the less likely you'll be able to rely on the results from your test. Strong covariates have the opposite effect: it can *increase* the power of your test.

General steps for ANCOVA

General steps are:

1. Run a regression between the independent and dependent variables.
2. Identify the residual values from the results.
3. Run an ANOVA on the residuals.

Assumptions for ANCOVA

Assumptions are basically the same as the ANOVA assumptions. Check that the following are true before running the test:

1. **Independent variables (minimum of two) should be categorical variables**.
2. **The dependent variable and covariate should be continuous variables** (measured on an interval scale or ratio scale.)
3. Make sure **observations are independent**. In other words, don't put people into more than one group.

Software can usually check the following assumptions.

1. **Normality**: the dependent variable should be roughly normal for each of category of independent variables.
2. Data should show **homogeneity of variance**.
3. **The covariate and dependent variable (at each level** of independent variable) should be **linearly related**.

4. Your data should be **homoscedastic** of Y for each value of X.
5. **The covariate and the independent variable shouldn't interact.** In other words, there should be homogeneity of regression slopes.