Reading: More about ANOVA

You've learned that analysis of variance—ANOVA—is a group of statistical techniques that test the difference of means between groups. ANOVA testing is useful when you want to test a hypothesis about group differences based on categorical independent variables. For example, if you wanted to determine whether changes in people's weight when following different diets are statistically significant or due to chance, you could use ANOVA to analyze the results. Data professionals routinely must ascertain if there are meaningful differences between groups in their data. This reading will examine more closely the intuition behind ANOVA using a worked example. Later in the program, you will learn how to implement ANOVA in Python.

An overview of ANOVA

The intuition behind ANOVA is to compare the variability between different groups with the variability within the groups. If they are comparable, then the differences between groups are more likely to be due to sampling variability. On the other hand, if the variability between groups is much larger than the variability expected from the samples within their respective groups, then those groups are probably drawn from significantly different subpopulations.

The variation between groups and within groups is calculated as sums of squares, which are then expressed as a ratio. This ratio is known as the F-statistic. The formula for each component of these calculations is presented in the worked example that follows.

Previously, you learned about one-way and two-way ANOVA. To review:

- **One-way ANOVA:** Compares the means of one continuous dependent variable based on three or more groups of **one** categorical variable
- **Two-way ANOVA:** Compares the means of one continuous dependent variable based on three or more groups of **two** categorical variables

To help you understand the intuition behind ANOVA, this reading will break down a worked example of a simple one-way ANOVA test.

One-way ANOVA

5 steps

There are five steps in performing a one-way ANOVA test:

1. Calculate group means and grand (overall) mean
2. Calculate the sum of squares between groups (SSB) and the sum of squares within groups (SSW)

3. Calculate mean squares for both SSB and SSW
4. Compute the F-statistic
5. Use the F-distribution and the F-statistic to get a p-value, which you use to decide whether to reject the null hypothesis

Example

Return to the example of students studying for an exam. Suppose that in this case you wanted to compare three different studying programs, A, B, and C to determine whether they have an effect on exam score. Here is the data:

Student	**Study program (X)**	**Exam score (Y)**
1	A	88
2	A	79
3	A	86
4	A	90
5	B	94
6	B	84
7	B	87
8	B	89
9	C	85
10	C	76
11	C	81
12	C	78

First, state your hypotheses:

H_0: $\mu_A = \mu_B = \mu_C$

The mean score of group A = the mean score of group B = the mean score of group C.

H_1: NOT ($\mu_A = \mu_B = \mu_C$)

The means of each group are not all equal. Remember, even if only one mean differs, that is sufficient evidence to reject the null hypothesis.

Next, determine your confidence level—the threshold above which you will reject the null hypothesis. This value is dependent on your situation and usually requires some domain knowledge. A common threshold is 95%.

Now, begin the steps of ANOVA.

Step 1
Calculate group means and grand mean. The grand mean is the overall mean of all samples in all groups.
The following table restructures the data in the previous table such that the scores for each study group are contained in their own column. Additionally, the mean score of each group has been calculated.

Study program A scores	Study program B scores	Study program C scores
88	94	85
79	84	76
86	87	81
90	89	78
Mean: 85.75	**Mean: 88.5**	**Mean: 80**

Grand mean (M_G) = 84.75

Step 2
A. Calculate the sum of squares between groups (SSB).

$$\text{SSB} = \sum_{g=1} n_g (M_g - M_G)^2$$

where:

n_g = the number of samples in the g^{th} group

M_g = mean of the g^{th} group

M_G = grand mean

→ **SSB** = $[4(85.75 - 84.75)^2 + 4(88.5 - 84.75)^2 + 4(80 - 84.75)^2]$
= 4 + 56.25 + 90.25
= 150.5

B. Calculate the sum of squares within groups (SSW).

$$\text{SSW} = \sum_{g=1} \sum_{i=1} (x_{gi} - M_g)^2$$

where:

x_{gi} = sample i of the g^{th} group

M_g = mean of the g^{th} group

The double summation acts like a nested loop. The outer loop is for each group and the inner loop is for all the samples in that group. So, for each sample in group 1, subtract from it the group's mean and square the result. Then, do the same thing with the samples in group 2, using the group 2 mean. Continue this way for all the groups and sum all the results.

The following table shows the squared difference between each observation and its group mean. It also contains the sums of these squared differences for each of the three study groups, A, B, and C.

Program A	Program B	Program C	$(x_{Ai} - M_A)^2$	$(x_{Bi} - M_B)^2$	$(x_{Ci} - M_C)^2$
88	94	85	5.06	30.25	25
79	84	76	45.56	20.25	16
86	87	81	0.06	2.25	1
90	89	78	18.06	0.25	4
M_A = 85.75	M_B = 88.5	M_C = 80	**Sum: 68.75**	**Sum: 53**	**Sum: 46**

→ **SSW** = 68.75 + 53 + 46

= **167.75**

Step 3
Calculate mean squares between groups and within groups. The mean square is the sum of squares divided by the degrees of freedom, respectively.

A. Mean squares between groups (MSSB):

$$\text{MSSB}= \frac{SSB}{k-1}$$

where:
k = the number of groups
Note: k–1 represents the degrees of freedom between groups

→ **MSSB** = $\frac{150.5}{3-1}$
= 75.25

B. Mean squares within groups (MSSW):

$$\text{MSSW}= \frac{SSW}{n-k}$$

where:
n = the total number of samples in all groups
k = the number of groups
Note: n–k represents the degrees of freedom within groups

→ **MSSW** = $\frac{167.75}{12-3}$
= 18.64

Step 4
Compute the F-statistic.
The F-statistic is the ratio of the mean sum of squares between groups (MSSB) to the mean sum of squares within groups (MSSW):

$$\text{F-statistic}= \frac{MSSB}{MSSW}$$

→ **F-statistic** = $\frac{75.25}{18.64}$
= 4.04

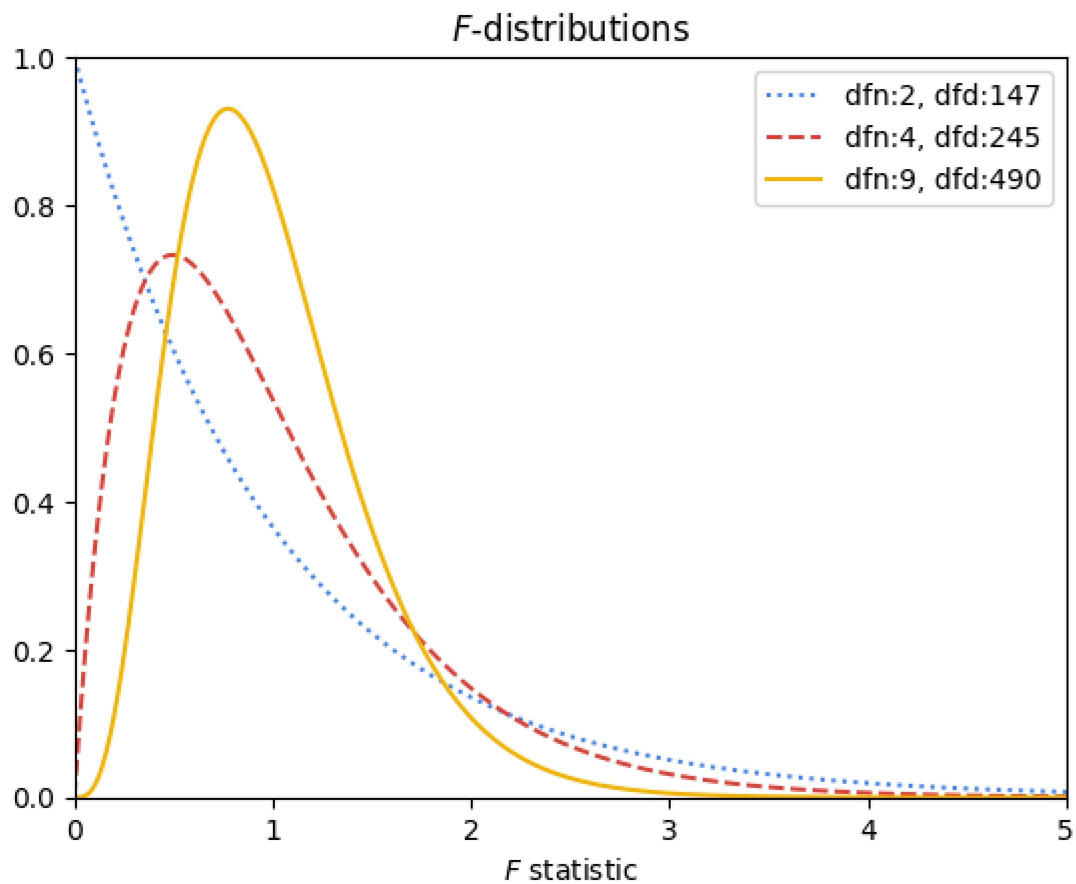
A higher F-statistic indicates a greater variability between group means relative to the variability within groups, suggesting that at least one group mean is significantly different from the others.

Step 5

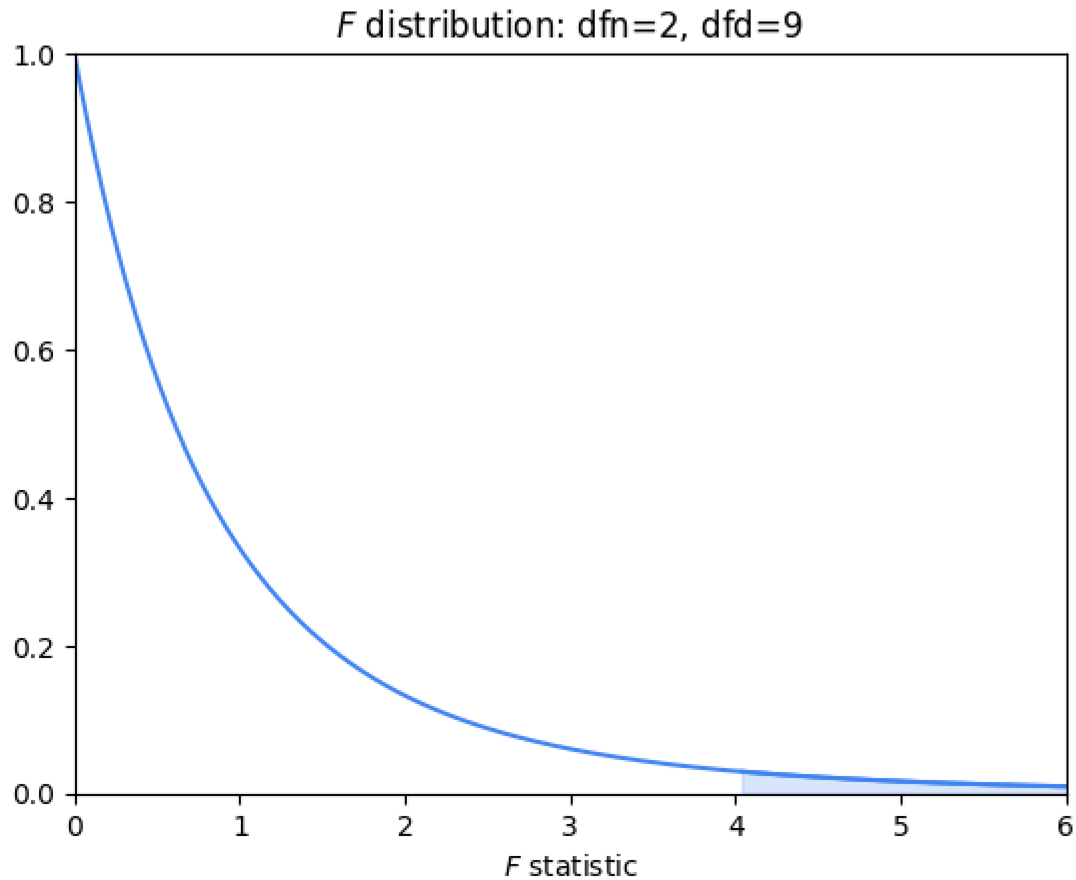
Use the F-distribution and the F-statistic to get a p-value, which you use to decide whether to reject the null hypothesis.

Similar to t-tests and X^2 tests, ANOVA testing finds the area under a particular probability distribution curve—the F-distribution—of the null hypothesis to determine a p-value. The larger the F-statistic, the lesser the area beneath the curve and the more evidence against the null hypothesis, thus resulting in a lower p-value.

The shape of the F-distribution curve is determined by the degrees of freedom between and within groups. Here is a graph depicting F-distributions for three, five, and 10 groups—each group containing 50 samples. Note that "dfn" represents the degrees of freedom in the numerator (between groups) and "dfd" represents the degrees of freedom in the denominator (within groups). Notice how the degrees of freedom affect the shape of the curve.



Similar to X^2 curves, F-distributions help determine the probability of falsely rejecting the null hypothesis. In the case of ANOVA, this probability is represented by the area of the F-distribution beneath the curve where x ≥ your F-statistic. For example, the following graph depicts the F-distribution for the exam scores example. It has two degrees of freedom in the numerator and nine degrees of freedom in the denominator. The area beneath the curve where x ≥ 4.04 (the computed F-statistic) is shaded.



You can use statistical software to calculate this area. You will learn how to do this in a later activity. In this case, the area beneath the F-distribution to the right of 4.04 is 0.05604. This is the probability of observing an F-statistic greater than 4.04 if the null hypothesis were true. Whether this is sufficient to reject the null hypothesis is a decision you make at the beginning of your hypothesis test. For example, if you decided that you wanted a confidence level of 95% or greater, you cannot reject the null hypothesis that the means of the distributions for each program of study are all the same, because the p-value is 0.056.

Assumptions of ANOVA

ANOVA will only work if the following assumptions are true:

1. The dependent values for each group come from normal distributions
    - Note that this assumption does NOT mean that *all* of the dependent values, taken together, must be normally distributed. Instead, it means that *within each group*, the dependent values should be normally distributed.
    - ANOVA is generally robust to violations of normality, especially when sample sizes are large or similar across groups, due to the central limit theorem. However, significant violations can lead to incorrect conclusions.
2. The variances across groups are equal

    - ANOVA compares means across groups and assumes that the variance around these means is the same for all groups. If the variances are unequal (i.e., heteroscedastic), it could lead to incorrect conclusions
3. Observations are independent of each other
    - ANOVA assumes that one observation does not influence or predict any other observation. If there is autocorrelation among the observations, the results of the ANOVA test could be biased.

Key takeaways

- ANOVA tests are statistical tests that examine whether or not the means of a continuous dependent variable are significantly different from one another based on the different levels of one or more independent categorical variables.
- It is sufficient for one group's mean to be significantly different from the others to reject the null hypothesis; however, ANOVA testing is limited in that it doesn't tell you *which* group is different. To make such a determination, other tests are necessary.
- ANOVA works by comparing the variance between each group to the variance within each group. The greater the ratio of variance between groups to variance within groups, the greater the likelihood of rejecting the null hypothesis.
- ANOVA depends on certain assumptions, so it is important to check that your data meets them in order to avoid drawing false conclusions. At the very least, if your data does not meet all of them, identify these violations.