# ANOVA

## Introduction

ANOVA (Analysis of Variance) is a method for generalizing statistical tests to multiple groups. As you'll see, ANOVA analyses the overall variance of a dataset by partitioning the total sum of squared deviations (from the mean) into the sum of squares for each of these groups and sum of squares for error. By comparing the statistical test for multiple groups, it can serve as a useful alternative to the t-tests you've encountered thus far when you wish to test multiple factors simultaneously.

## Objectives

You will be able to:

- Explain the methodology behind ANOVA tests
- Use ANOVA for testing multiple pairwise comparisons

## Explanation of ANOVA

To perform ANOVA, you begin with sample observations from multiple groups. Since ANOVA is looking to explain the total variance as combinations of variance from the various groups, you typically design a multiple groups experiment to test various independent factors that we hypothesize may influence the overall result. For example, imagine an email campaign designed to optimize donation contributions. In order to get the most money in donations, one might send out two different emails, both copies being identical except for the subject line. This would form a sensible hypothesis test, but if you wanted to test multiple changes simultaneously, swapping out subject line, time sent, thank you gift offers, or other details in the email campaign, then ANOVA would be a more appropriate methodology. In this scenario, you would change one or more of these various features and record the various donations. Once you have sample observations from various combinations of these features, you can then use ANOVA to analyze and compare the effectiveness of the individual features themselves.

The general idea is to break the sum of squared deviations into multiple parts: the sum of squared deviations of the mean of each of the test groups to the observations within the group itself, and the sum of squared deviations of the mean of these test groups to the mean of all observations.

This is easier to understand through the context of an example. For the email case described above, ANOVA would compare the variance of donations within each of the groups to the overall variance of all donations (or lack thereof) as a whole. If the variance of a single group's donations versus that of the overall sample is substantial, there is reason to reject the null hypothesis for that feature. This forms the foundation of the f-test which is at the heart of ANOVA.

Recall that you would not perform multiple t-tests with such a scenario because of the multiple comparisons problem. Type I errors will be confounded when conducting multiple t-tests. While the alpha threshold for any one test might be 0.05, it would not be surprising to reject the null hypothesis in at least one of these cases, just by pure chance, if you conduct 5 or 10 t-tests.

## ANOVA in Python

In [1]:

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

## Loading the data

As usual, we start by loading in a dataset of our sample observations. This particular table is of salaries in IT and has 4 columns:

- S - the individuals salary
- X - years of experience
- E - education level (1-Bachelors, 2-Masters, 3-PHD)
- M - management (0-no management, 1-yes management)

In [2]:

```
df = pd.read_csv('IT_salaries.csv')
```

```
df.head()
```

Out[2]:

|   | S | X | E | M |
|---|---|---|---|---|
| 0 | 13876 | 1 | 1 | 1 |
| 1 | 11608 | 1 | 3 | 0 |
| 2 | 18701 | 1 | 3 | 1 |
| 3 | 11283 | 1 | 2 | 0 |
| 4 | 11767 | 1 | 3 | 0 |

## Generating the ANOVA table

In order to generate the ANOVA table, you first fit a linear model and then generate the table from this object. Our formula will be written as:

`Control_Column ~ C(factor_col1) + factor_col2 + C(factor_col3) + ... + X`

*We indicate categorical variables by wrapping them with* `C()` .

In [3]:

```
formula = 'S ~ C(E) + C(M) + X'
lm = ols(formula, df).fit()
table = sm.stats.anova_lm(lm, typ=2)
print(table)
```

```
                sum_sq    df           F        PR(>F)
C(E)      9.152624e+07   2.0   43.351589  7.672450e-11
C(M)      5.075724e+08   1.0  480.825394  2.901444e-24
X         3.380979e+08   1.0  320.281524  5.546313e-21
Residual  4.328072e+07  41.0         NaN           NaN
```

## Interpreting the table

For now, simply focus on the outermost columns. On the left, you can see our various groups, and on the right, the probability that the factor is indeed influential. Values less than 0.05 (or whatever we set $\alpha$ to) indicate rejection of the null hypothesis. In this case, notice that all three factors appear influential, with management being the potentially most significant, followed by years experience, and finally, educational degree.

## Summary

In this lesson, you examined the ANOVA technique to generalize testing methods to multiple groups and factors.