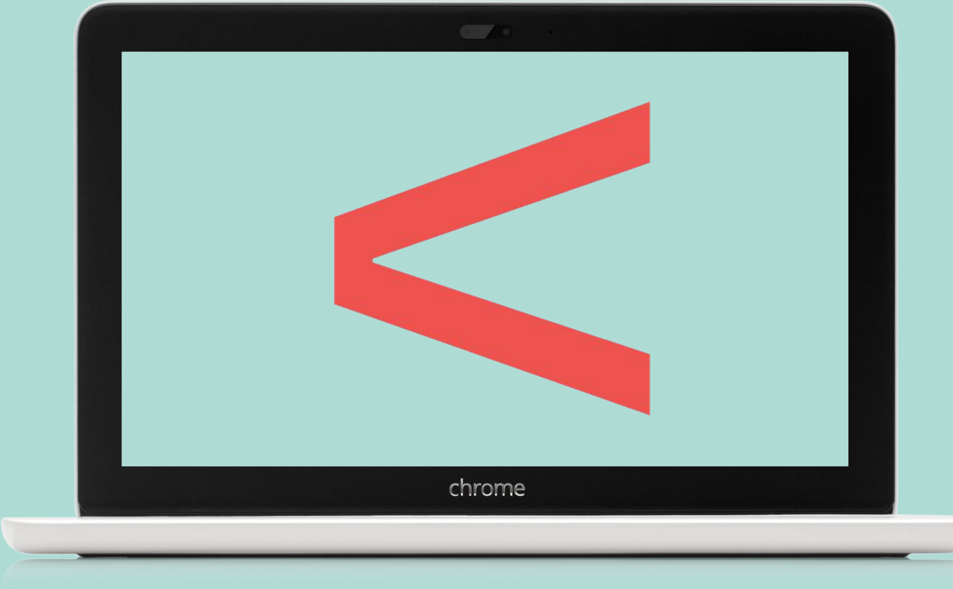
Introduction to ANOVA



ANOVA
What is it?

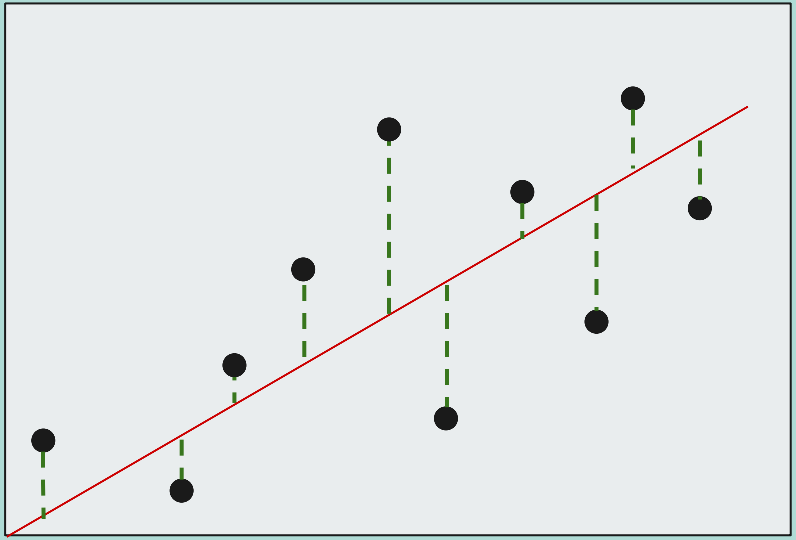
ANOVA stands for Analysis of Variance. It is a test that compares if there is **a difference in the means (averages) of a condition between more than two independent comparison groups**.

Usually, it uses an independent categorical variable to predict a dependent continuous variable.

What is variance?

The variance quantifies how much each number in a dataset varies from the mean of the dataset.

If the variance is small, it means the data points tend to be close to the mean, and if the variance is large, the data points are spread out over a larger range of values.



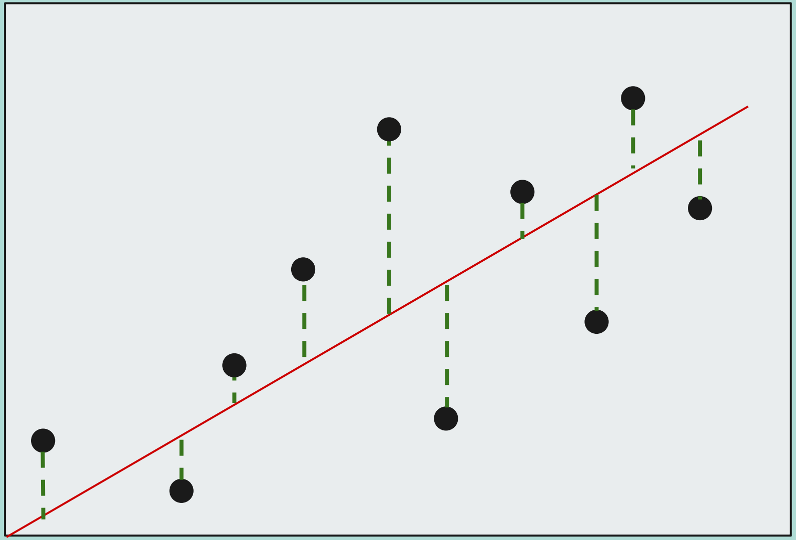
How to calculate the variance?

1. Find the mean of the dataset.
2. Subtract the mean from each data point to find the deviation of each point from the mean.
3. Square each of these deviations. This step ensures that all deviations are positive and gives more weight to larger deviations.
4. Sum up all the squared deviations.
5. Divide this sum by n - 1 (where n is the number of data points in the sample).

What is standard deviation?

It is the square root of the variance. It measures the average distance between each data point and the mean.

A low standard deviation indicates that the values in a dataset are close to the mean, while a high standard deviation indicates that the values are spread out over a wider range.



Types of ANOVA

One-way ANOVA: Used when there's one independent variable. For instance, testing the effect of three different diets on weight loss.

Two-way ANOVA: Used when there are two independent variables. For instance, testing the effect of diet type and exercise regimen on weight loss.

Repeated measures ANOVA: Used when the same subjects are used for each treatment (e.g., a medical study in which patients are compared before and after a treatment).

One-way ANOVA

Null Hypothesis: The means in the groups are equal. (No variation between groups.)

Alternative Hypothesis: At least one of the group mean is different. (There is variation between groups.)

Relationship between one variable X that is categorical (divided into at least 3 groups) with a Y variable (numeric) is statistically significant.

One-way ANOVA example

	LABEL		
	low	medium	high
ALCOHOL % VOL.	12.0	8.8	12.8
	9.7	9.5	12.8
	10.8	10.1	10.5
	9.7	9.9	10.7
	9.5	9.9	10.7
	9.3	8.8	12.1
	10.2	9.5	12.1
	12.8	10.1	12.7
	10.0	9.6	9.6
	8.6	11.0	12.6
Individual average:	10.26	9.72	11.66

Grand average: 10.55

One-way ANOVA example in Python.

```
from scipy import stats

# Perform ANOVA test
F, p = stats.f_oneway(
    red_wine[red_wine['quality_label'] == 'low']['alcohol'],
    red_wine[red_wine['quality_label'] == 'medium']['alcohol'],
    red_wine[red_wine['quality_label'] == 'high']['alcohol']
)

# Print results
print('ANOVA test for mean alcohol levels across wine samples with
different quality ratings')
print('F Statistic:', F)
print('p-value:', p)
```

One-way ANOVA example in Python.

F-Statistic (F-value): The F-statistic is a ratio of two variances. In the context of ANOVA, it represents the ratio of the variance between the group means to the variance within the groups.

P-Value: The p-value is a measure of evidence against the null hypothesis. A smaller p-value suggests stronger evidence against the null hypothesis. If the p-value is below a predetermined significance level (alpha), you might reject the null hypothesis.

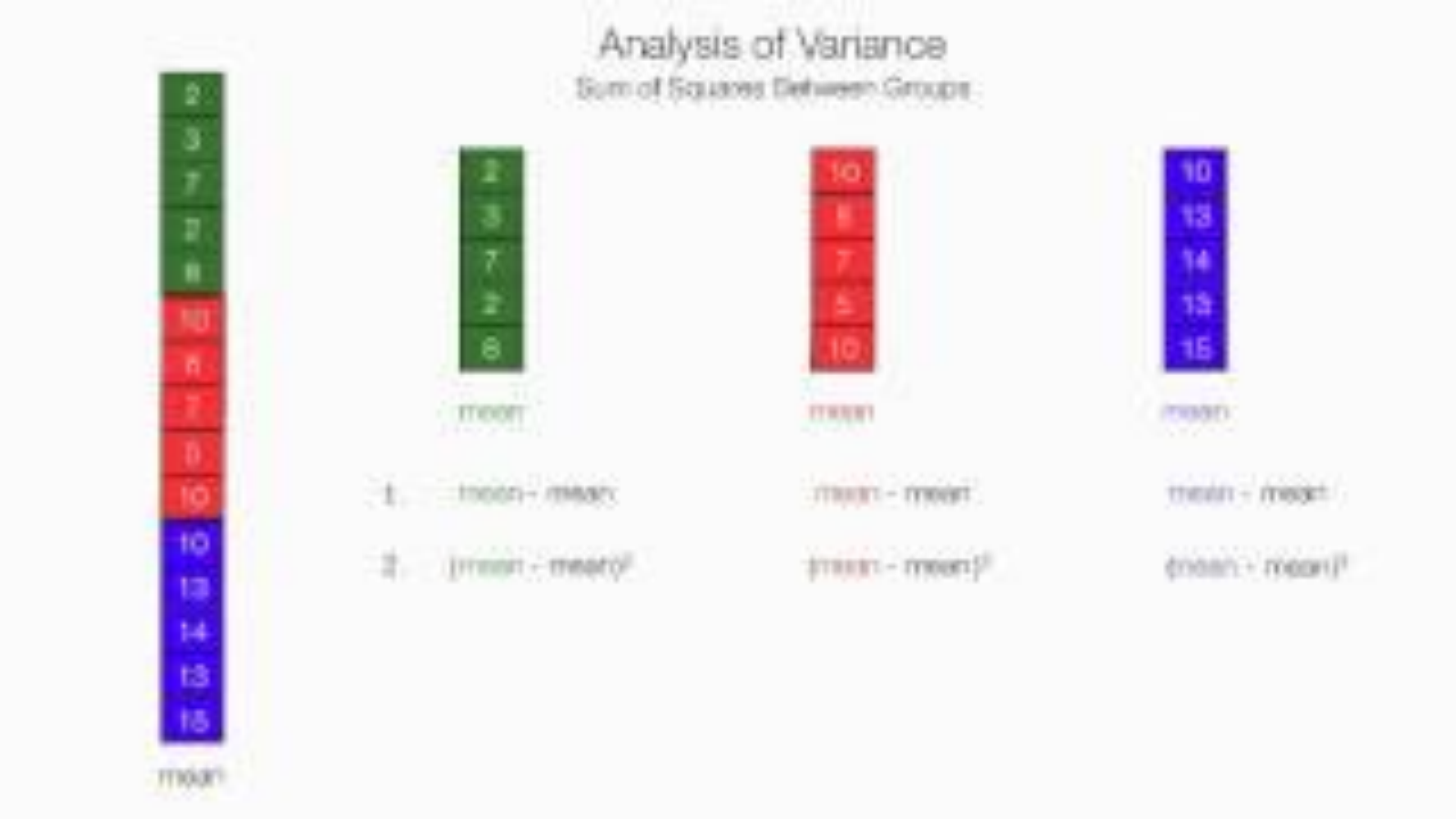
Alpha (α): Alpha, also known as the significance level, is a predetermined threshold that you choose before conducting a hypothesis test. A common choice for alpha is 0.05 (5%). Please remember that **α** = 0.05 criterion is a convention, not an absolute standard. Depending on the field of study or the specific research question, other thresholds like **α** = 0.01 or **α** = 0.10 might be used.

https://www.statology.org/anova-f-value-p-value/
https://towardsdatascience.com/anova-test-with-python-cfbf4013328b

Limitations

- **Data Shape Matters:** ANOVA assumes that the groups you're comparing have roughly equal variability and follow a normal distribution. If your data is very different from this assumption, ANOVA might not give accurate results.
- **Group Sizes and Outliers:** ANOVA works best when group sizes are similar and there aren't extreme outliers. Uneven group sizes and outliers can make the results less reliable.
- **Picking Important Differences:** ANOVA can show if groups are different, but not which groups are different from each other. Extra tests are needed for that, but these tests can sometimes lead to false results if you're not careful.
- **Don't Overdo It:** If you run ANOVA multiple times on the same data, the chance of making mistakes goes up. Make sure to adjust your approach if you're comparing lots of things.

ANOVA is a useful tool, but it has some rules you need to follow for it to work well and give you meaningful insights.



Analysis of Variance

Sum of Squares Between Groups

1. mean - mean mean - mean mean - mean

2. $(\text{mean} - \text{mean})^2$ $(\text{mean} - \text{mean})^2$ $(\text{mean} - \text{mean})^2$