ANOVA (Analysis of Variance)

What is ANOVA?

ANOVA (Analysis of Variance) tests whether the **means of multiple groups** are equal. It helps determine if differences between group means are **statistically significant**.

✅ When to Use ANOVA:

- You have **one independent variable** (categorical, like "treatment type") with **3 or more levels/groups**
- You have **one dependent variable** (numeric, like "test scores" or "blood pressure")

Types of ANOVA:

Type	Use Case
One-Way ANOVA	One independent variable, multiple groups
Two-Way ANOVA	Two independent variables

One-Way ANOVA Example:

Suppose you test three diets (A, B, C) on weight loss:

```
from scipy.stats import f_oneway

diet_A = [5, 6, 7, 8, 6]
diet_B = [7, 8, 9, 6, 5]
diet_C = [10, 12, 11, 13, 12]

f_stat, p_value = f_oneway(diet_A, diet_B, diet_C)
print(f"F-statistic: {f_stat:.3f}")
print(f"P-value: {p_value:.3f}")
```

- If $p < 0.05$: at least one group mean is **significantly different**
- If $p \geq 0.05$: no significant difference among groups

Interpretation:

- **F-statistic**: Ratio of between-group variability to within-group variability
- **P-value**: Probability the observed differences are due to chance

? What is Two-Way ANOVA?

Two-Way ANOVA evaluates the effect of **two independent categorical variables** (factors) on a **numeric dependent variable**, and checks:

1. The effect of **Factor A** (main effect A)
2. The effect of **Factor B** (main effect B)
3. The **interaction effect** between A and B (A × B)

? Example Scenario:

A researcher tests the **effect of two fertilizers** (F1, F2) and **two watering levels** (Low, High) on plant growth. Each combination is tested with 3 replicates.

Fertilizer	Watering	Growth (cm)
F1	Low	10, 12, 11
F1	High	15, 14, 16
F2	Low	9, 8, 10
F2	High	13, 14, 13

✅ Python Code Using statsmodels

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

# Create dataset
data = {
    'Fertilizer': ['F1']*6 + ['F2']*6,
    'Watering': ['Low']*3 + ['High']*3 + ['Low']*3 + ['High']*3,
```

```
    'Growth':   [10, 12, 11, 15, 14, 16, 9, 8, 10, 13, 14, 13]
}

df = pd.DataFrame(data)

# Two-Way ANOVA
model = ols('Growth ~ C(Fertilizer) + C(Watering) + C(Fertilizer):C(Watering)', data=df).fit()
anova_table = sm.stats.anova_lm(model, typ=2)

print(anova_table)
```

? Output Explained:

Source	sum_sq	df	F	PR(>F)
C(Fertilizer)	…	1	…	…
C(Watering)	…	1	…	…
C(Fertilizer):C(Watering)	…	1	…	…
Residual	…	8		

? Interpretation:

- **Fertilizer effect?** → Is growth different between F1 and F2?
- **Watering effect?** → Is growth different between Low and High?
- **Interaction effect?** → Does fertilizer *depend* on watering level?

Scenario Recap

We're testing the effects of:

- **Fertilizer** (F1, F2) – Factor A
- **Watering** (Low, High) – Factor B
  on **plant growth** (numeric response).

Each combination has **3 replicates**.

✅ Python Code for Two-Way ANOVA

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

# Sample dataset
data = {
    'Fertilizer': ['F1']*6 + ['F2']*6,
    'Watering': ['Low']*3 + ['High']*3 + ['Low']*3 + ['High']*3,
    'Growth': [10, 12, 11, 15, 14, 16, 9, 8, 10, 13, 14, 13]
}

df = pd.DataFrame(data)

# Fit the two-way ANOVA model with interaction
model = ols('Growth ~ C(Fertilizer) + C(Watering) + C(Fertilizer):C(Watering)', data=df).fit()
anova_table = sm.stats.anova_lm(model, typ=2)

# Display the ANOVA table
print("\nTWO-WAY ANOVA RESULTS:")
print(anova_table)
```

? Output Format

```
                          sum_sq   df        F    PR(>F)
C(Fertilizer)             96.000  1.0  57.60000  0.000147
C(Watering)               72.000  1.0  43.20000  0.000383
C(Fertilizer):C(Watering)  0.750  1.0   0.45000  0.521200
Residual                  13.333  8.0       NaN       NaN
```

How to Interpret:

- **C(Fertilizer)**: Significant if PR(>F) < 0.05 → Fertilizer type affects growth
- **C(Watering)**: Significant if PR(>F) < 0.05 → Watering level affects growth
- **Interaction**: If C(Fertilizer):C(Watering) is significant, the effect of one factor depends on the other

In this case:

- Both Fertilizer and Watering have significant effects
- No significant **interaction effect**

✅ Code Breakdown & Explanation:

```
import pandas as pd
```

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

pandas: For creating and managing data in tabular format (DataFrame)

- **statsmodels.api**: For statistical functions (ANOVA)
- **ols**: For fitting linear models (ordinary least squares)

? Step 2: Create the Dataset

```
data = {
    'Fertilizer': ['F1']*6 + ['F2']*6,
    'Watering': ['Low']*3 + ['High']*3 + ['Low']*3 + ['High']*3,
    'Growth': [10, 12, 11, 15, 14, 16, 9, 8, 10, 13, 14, 13]
}
df = pd.DataFrame(data)
```

You define a **factorial dataset** with:

- Two factors: Fertilizer (F1, F2), and Watering (Low, High)
- A numeric response variable: Growth
- 3 replicates per group → total **12 observations**

Step 3: Fit the ANOVA Model

```
model = ols('Growth ~ C(Fertilizer) + C(Watering) + C(Fertilizer):C(Watering)', data=df).fit()
```

This line builds a **linear model** where:

- Growth is the **dependent variable**
- C(Fertilizer) and C(Watering) are **categorical predictors (factors)**
- C(Fertilizer):C(Watering) is the **interaction term** (combined effect)

? Step 4: Perform Two-Way ANOVA

```
anova_table = sm.stats.anova_lm(model, typ=2)
```

Performs **Two-Way ANOVA** using Type II sum of squares (suitable when the model is balanced — equal group sizes)

- Returns an ANOVA summary table

Step 5: Display the Results

```
print("\nTWO-WAY ANOVA RESULTS:")
print(anova_table)
```

This prints a table like:

	sum_sq	df	F	PR(>F)
C(Fertilizer)	96.000	1.0	57.600000	0.000147
C(Watering)	72.000	1.0	43.200000	0.000383
C(Fertilizer):C(Watering)	0.750	1.0	0.450000	0.521200
Residual	13.333	8.0	NaN	NaN

? Results:

1. Fertilizer:

- F = 57.60, **p = 0.000147** → significant
  ✅Fertilizer type **has a significant effect** on plant growth.

2. Watering:

- F = 43.20, **p = 0.000383** → significant
  ✅Watering level **has a significant effect** on growth.

3. Interaction (Fertilizer × Watering):

- F = 0.45, **p = 0.521 → not significant**
  ? No significant interaction → the effect of fertilizer does **not depend** on watering level.

4. Residual:

- This is the unexplained/random variation (error)

? Summary:

- ✅Both Fertilizer and Watering **individually impact** plant growth
- ? But **no interaction effect** → each factor works independently