The Core Difference between univariate and multivariate outlier detection methods

- Univariate Approach: Looks at one variable at a time. It asks: "Is this data point extreme compared to the rest of the data in this single column?"

- Multivariate Approach: Looks at multiple variables simultaneously. It asks: "Is this combination of values across several columns unusual, even if each individual value seems normal on its own?"

IQR / Box Plot

What is it?

The Interquartile Range (IQR) method, often visualized with a Box Plot (or Whisker Plot), is a simple, non-parametric technique for identifying outliers. It doesn't assume a specific statistical distribution (like normality), making it very robust and widely applicable.

Core Concept

The IQR measures the statistical "spread" of the middle 50% of your data. It is calculated as the difference between the 3rd Quartile (Q3) and the 1st Quartile (Q1).

- **Q1 (First Quartile):** The median of the lower half of the dataset. 25% of the data lies below this point.
- **Q3 (Third Quartile):** The median of the upper half of the dataset. 75% of the data lies below this point.
- **IQR = Q3 - Q1**

How to Find Outliers: "Tukey's Fences"

Outliers are defined as data points that fall significantly outside the "fences" built around the central data. The standard fences are:

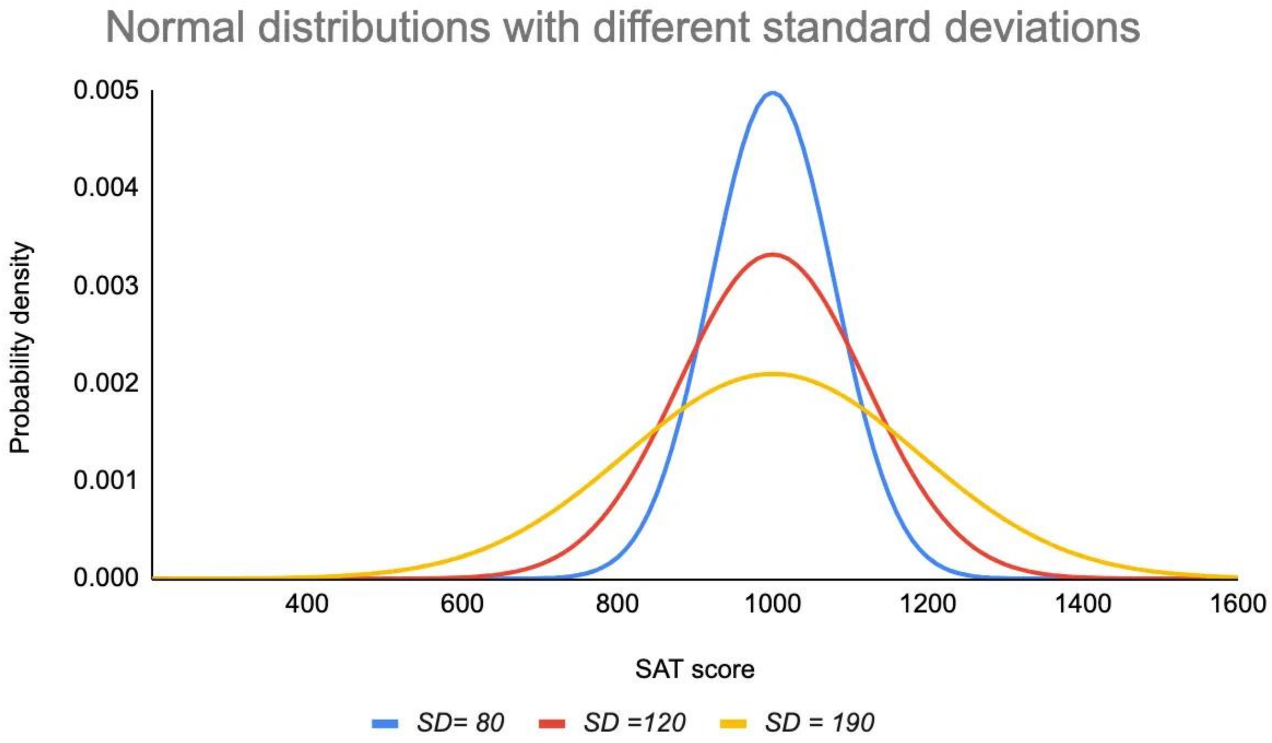
- **Lower Fence = Q1 - (1.5 * IQR)**
- **Upper Fence = Q3 + (1.5 * IQR)**

1. Any data point **below the Lower Fence** or **above the Upper Fence** is considered an outlier.
2. For identifying **extreme outliers**, a stricter multiplier of **3 * IQR** can be used.

Z-Score

What is it?

The Z-Score (or Standard Score) is a parametric statistical method that measures how many **standard deviations** a single data point is away from the **mean** of the dataset. It is best used when data is roughly **normally distributed** (following a bell curve).



Normal distributions with different standard deviations

Core Concept

The standard deviation (σ) measures the average amount of variation or dispersion in a dataset.

A low standard deviation means data points are clustered closely around the mean.

A high standard deviation means data points are spread out over a wider range.

3. The Z-Score Formula

The Z-Score for a single data point is calculated as follows:

$Z = (X - \mu) / \sigma$

X = The individual data point

μ (mu) = The mean (average) of the dataset

σ (sigma) = The standard deviation of the dataset

4. How to Identify Outliers

After calculating the Z-Score for every point, you apply a threshold. A common threshold is:

Any data point with |Z-Score| > 3 is considered an outlier.
(This means the point is more than 3 standard deviations above or below the mean).

A more conservative threshold of |Z-Score| > 2 can be used to flag more potential outliers.

What is PCA?

Principal Component Analysis (PCA) has been widely used in various research fields (e.g., electromyography, EMG) to reduce the dimensionality of the original sensor space and simplify subsequent analyses. By means of an orthogonal rotation, PCA linearly transforms a set of input data channels into an equal number of linearly-uncorrelated variables (Principal Components, PCs) that each successively account for the largest possible portion of remaining data variance (Kambhatla and Leen, 1997).

WHY PCA?

- Reduce data complexity
  - By identifying linear combinations of EEG channels or features that maximize variance
- Correct artifacts
- Enhance analysis
- Cost of Computation

TYPES OF PCA IN EEG STUDIES

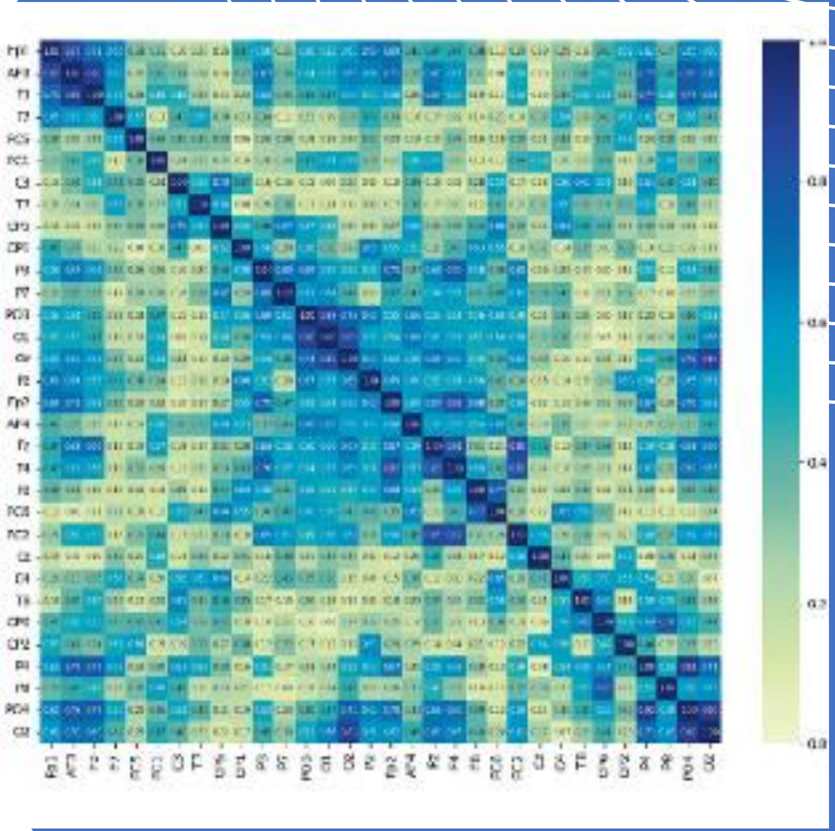
Type of PCA	Description	Application/Feature
Standard PCA	Identifies principal components (linearly uncorrelated orthogonal variables) to reduce dimensionality.	Used for feature extraction, noise reduction, and pattern recognition in EEG epochs
PCA for Channel	Modified PCA designed to Find Channel in EEG signals.	Enhances EEG for channel re
PCA with Wavelet Transform	Combines PCA and wavelet transform for hybrid dimensionality reduction.	Extracts features from EEG signals while preserving temporal-frequency information
PCA in Frequency Domain	Applies PCA directly in the frequency domain instead of the time domain.	Separates multichannel EEG epochs into spatially and temporally independent components
PCA for Preprocessing	Integrates PCA as part of semi-automatic preprocessing pipelines. Modified PCA designed to handle outliers or noise in EEG signals.	Reduces noise and redundancy before further analysis (e.g., ICA or neural networks) Enhances robustness in EEG classification models by improving resilience to artifacts

Covariance Matrix

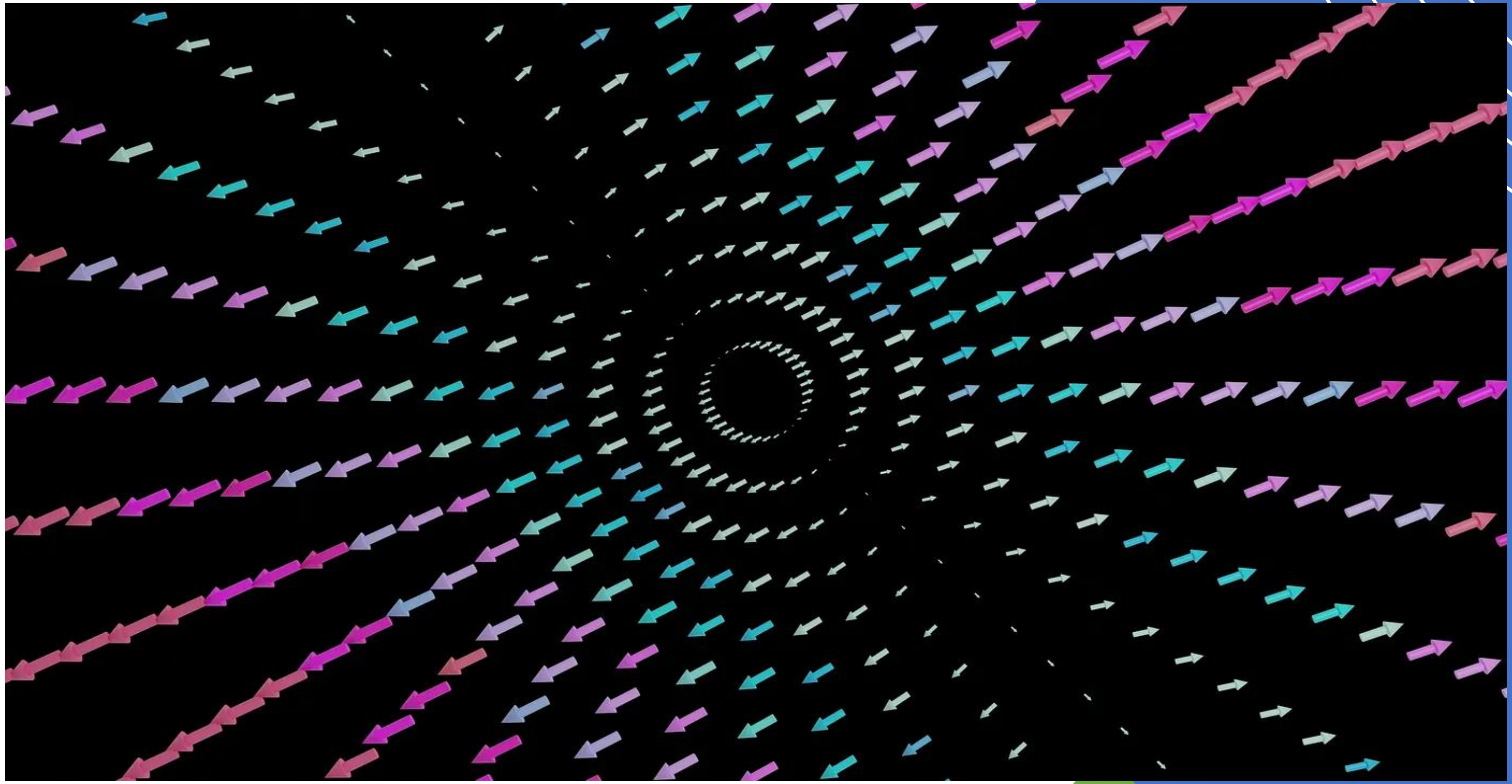
C = [Cov(X, X) Cov(X, Y) Cov(X, Z)]
[Cov(Y, X) Cov(Y, Y) Cov(Y, Z)]
[Cov(Z, X) Cov(Z, Y) Cov(Z, Z)]

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

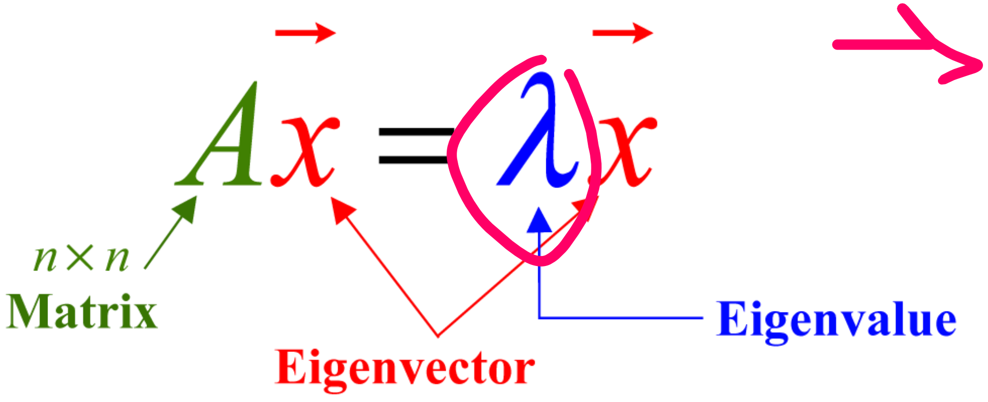
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$



Eigenvectors & Eigenvalues



Eigenvectors & Eigenvalues



$$A\vec{x} = \lambda\vec{x}$$

$n \times n$

Matrix

Eigenvector

Eigenvalue

PCA & PIZZA

Clustering Approach

How Clustering Identifies Outliers

The core principle is that normal data points belong to tight, dense clusters, while outliers do not. There are two main ways clustering is used for this purpose:

As a Direct Byproduct: Some algorithms, like DBSCAN, have a built-in concept of outliers ("noise points").

As a Post-Clustering Analysis: For algorithms that assign every point to a cluster (e.g., K-Means), we can analyze the results to find unusual points.

How to deal with outliers?
First, Check it again!!

Heat Index:
178°F (81°C)

Rafsanjan
Kerman
Yasuj
Shiraz
Sirjan
Bam
Zahedan
Khash
Jiroft
Iranshahr
Qeshm Dayrestan International Airport
Persian Gulf
Bandar Abbas
Minab
Dubai
Ras Al-Khaimah
Qatar
Fujairah
Sohar
Muscat
Abu Dhabi
Al Ain
Nizwa
Sur

How to deal with outliers?

Always investigate an outlier before deciding how to handle it…

An outlier could be:

1. A data entry error (e.g., a misplaced decimal, 300 years old instead of 30).
2. A measurement error (e.g., a sensor malfunction).
3. A natural variation in the data (e.g., the actual net worth of Jeff Bezos in a dataset of average people).
4. A novel Discovery.

How to deal with outliers?

1. Trimming (or Removal)
This method involves completely removing the outlier data points from the dataset.

How it works: You use a method to define outliers (e.g., data points beyond ±3 standard deviations, or outside the 1.5*IQR range) and simply delete those rows.

When to use:

- You are sure the outliers are due to data entry or measurement errors and cannot be corrected.
- The outliers are not representative of the population you are studying.
- You have a large enough dataset that losing a few rows won't impact the analysis.

ADV: Simple, clean, effective at eliminating the influence of the outlier.

DADV: You lose information (data points). If the outliers are real and important, you are removing valuable information.

How to deal with outliers?

2. Capping (or Winsorizing)

This method involves limiting the extreme values by replacing them with a certain percentile value.

How it works:

For example, you can cap all values above the 99th percentile to the value of the 99th percentile.

Similarly, you cap all values below the 1st percentile to the value of the 1st percentile.

A common method is using the Interquartile Range (IQR):

Upper Bound = Q3 + (1.5 * IQR)

Lower Bound = Q1 - (1.5 * IQR)

Any value above the Upper Bound is set to the Upper Bound value. Any value below the Lower Bound is set to the Lower Bound value.

How to deal with outliers?

When to use:

When you want to retain all data points in your dataset.

When the outliers are likely to be errors or non-representative, but you don't want to lose the entire row for other variables.

ADV: Doesn't reduce the sample size. Less disruptive than trimming.

DADV: Distorts the distribution and variances by compressing the extreme values. The choice of the percentile is arbitrary.

How to deal with outliers?

3. Discretization (Binning)
This is the process of converting a continuous variable into a categorical variable by creating "bins" or intervals.

How it works: You transform the numerical value into a category.

Example: Instead of using "Age" as a number, you bin it into categories like [0-18], [19-35], [36-60], [60+]. An outlier age of 200 would simply fall into the [60+] bin.

When to use:

When the exact value of the data is less important than the range it falls into.

For algorithms that work better with categorical data.

ADV: Robust to outliers as they are grouped into a top or bottom category.

DADV: You lose information about the actual value and the differences within a bin.

How to deal with outliers?

4. Transformation
Applying a mathematical transformation to the data to reduce the impact of outliers by compressing the scale.

Common Transformations:

Log Transformation: log(x) is very effective for right-skewed data. It brings in large values closer to the rest.
Square Root Transformation: sqrt(x)
Box-Cox Transformation: A more sophisticated, parameterized transformation that finds the best power transformation to make the data look more normal.

When to use: For highly skewed data and non-linear models. Essential for many **linear models** that assume normality.

ADV: Can make the data more "normal" (bell-shaped), which is a requirement for many statistical techniques.

DADV: Makes the data harder to interpret. The model coefficients will be in the transformed scale.

How to deal with outliers?

5. Treating them as a Separate Category

If you suspect the outliers represent a different phenomenon, you can flag them.

How it works: Create a new binary feature (column) called it **outlier**. For the outlier points, you set this to 1 and for non-outliers to 0. You can then cap or transform the original numerical value.

When to use: When you believe outliers represent a different group or state (e.g., "faulty sensor" vs. "normal sensor").

ADV: Captures the information that an outlier occurred without letting its extreme value distort the analysis.

How to deal with outliers?

6. Using Robust Algorithms

Choose models that are inherently less sensitive to outliers.

Robust Models: Tree-based models (Random Forest, Gradient Boosting, Decision Trees) are generally robust to outliers. They split data based on values, and an outlier will just end up in its own node.

Non-Robust Models: Linear Regression, Logistic Regression, and models that rely on distance-based calculations (like K-Means clustering or K-Nearest Neighbors) are highly sensitive to outliers.

When to use: When you have limited time for preprocessing or when the presence of extreme values is a natural part of your data.

How to deal with outliers?

Scenario	**Recommended Method**
The outlier is a clear error and you have plenty of data.	Trimming (Removal)
The outlier is an error but you want to keep the row.	Capping (Winsorizing)
The data is highly skewed.	Transformation (e.g., Log)
The outlier represents a different state.	Treat as a Separate Category
You're building a model and want simplicity.	Use Robust Algorithms (Random Forest)
The exact value is less important than its range.	Discretization (Binning)

Comparisons of Means

A direct comparison of laboratory and community EEG recordings for neurodevelopmental research

Comparison of Means in EEG Research

Example: Resting-State Alpha Power in Frontal Regions

Research Context:
Comparison of mean alpha power (8-12 Hz) between children with autism spectrum disorder (ASD) and typically developing (TD) children during resting-state EEG recordings 1.

Hypotheses:

H_0 (Null Hypothesis): The mean alpha power in frontal regions is equal in both groups ($\mu_1 = \mu_2$).

H_1 (Alternative Hypothesis): The mean alpha power in frontal regions differs between groups ($\mu_1 \neq \mu_2$).

Statistical Test:
Independent samples *t*-test (parametric test assuming normality and equal variances).

Results:

ASD Group: Mean alpha power = 4.2 μV^2

TD Group: Mean alpha power = 6.1 μV^2

p-value = 0.003

Interpretation:
The p-value (< 0.05) indicates rejection of the null hypothesis. Children with ASD show **significantly reduced alpha power** in frontal regions compared to TD children, suggesting altered neural oscillations potentially linked to differences in attention or cortical inhibition

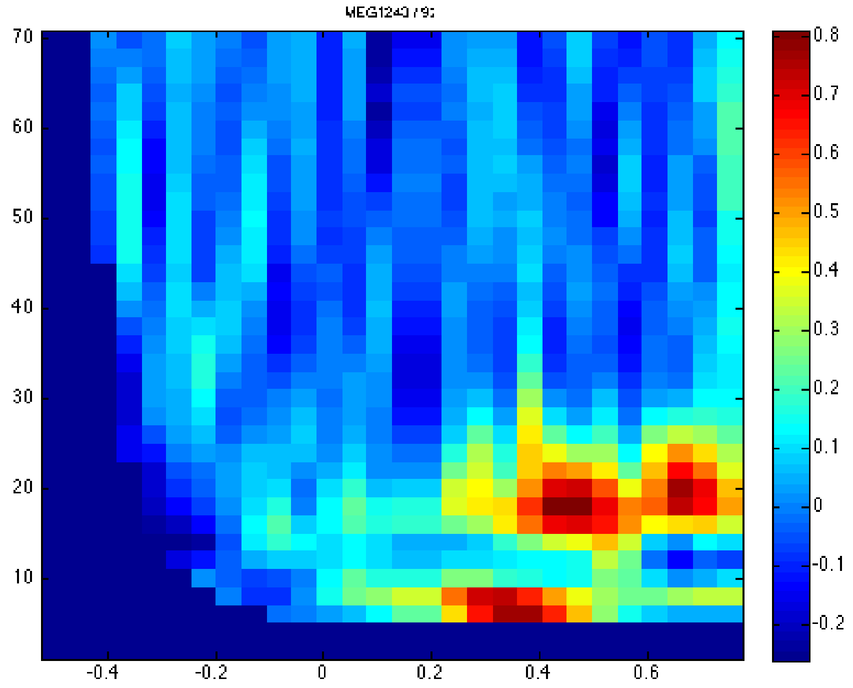
https://www.nature.com/articles/s41598-025-03569-5

Choosing the Right Statistical Test for EEG Data

Core Principle: Your choice depends on two factors:

Number of Groups: Are you comparing 2 groups (e.g., Patients vs. Controls) or more than 2 groups (e.g., Mild, Moderate, Severe ASD)?

Number of Time Points or Conditions: Are you measuring at a single time point, or repeatedly (e.g., Pre-Test vs. Post-Test, or across multiple conditions like Rest, Task1, Task2)?



Scenario 1: Two Groups, One Time Point/Condition

Research Question: **Do patients with Major Depressive Disorder (MDD) have different frontal alpha asymmetry scores than healthy controls at rest?**

Data Structure:

Between-Subjects Factor: Group (MDD vs. Control). Each participant is in only one group.

Dependent Variable (DV): A single EEG value per subject (e.g., mean frontal alpha asymmetry index).

Recommended Test: Independent Samples t-test

Example:

Comparison of Two Independent Groups

Hypotheses:

H_0: Mean alpha asymmetry is equal between MDD and control groups.

H_1: Mean alpha asymmetry differs between groups.

Statistical Test: Independent samples *t*-test.

Results:

MDD Group: Mean Asymmetry = -0.08

Control Group: Mean Asymmetry = +0.05

p-value = 0.02, Cohen's *d* = 0.7

Interpretation:

We reject the null hypothesis. The MDD group shows significantly greater relative right-frontal activity (a negative asymmetry index) compared to controls, consistent with models of withdrawal-related motivation.

Scenario 2: One Group, Two or More Time Points/Conditions

This is used when you measure the same group of people under different conditions or at different times.

Research Question: Does a mindfulness intervention alter theta power in our participants? We measure theta power during a meditation task before and after an 8-week course.

Data Structure:

Within-Subjects Factor: Time (Pre-Intervention vs. Post-Intervention). Each participant is measured twice.

Dependent Variable (DV): Theta power at each time point.

Recommended Test: Paired Samples t-test (for 2 time points) or Repeated Measures ANOVA (for 3+ time points, e.g., Pre, Mid, Post).