What is a Violin Plot?

A **violin plot** combines the features of a box plot and a kernel density plot to provide a richer visualization of data distribution. While box plots summarize central tendency and variability, violin plots also show the **probability density**, capturing the shape, peaks, and potential multimodality of the data. This makes violin plots especially useful for revealing nuanced distribution patterns.

Notes on Excel Implementation

- **Data Input**: Users select a data range in Excel (minimum 3 rows × 1 column). Columns represent variables; rows represent samples.
- **Header Detection**: If the first row contains non-numeric values, it is treated as a header (optional).
- **Kernel Density Estimation (KDE)**: A Gaussian KDE is computed for each variable.
- **Bandwidth Selection**: The bandwidth for KDE is determined using **Scott's Rule**.
- **Plot Construction**: The KDE curve is mirrored horizontally around a fixed offset (typically 1) to form the violin shape.
- **Summary Statistics**: The **median** and **mean** are calculated for each distribution and shown on the plot.

Formula Used in Violin Plot

Bandwidth Selection (Scott's Rule):

$$h = \sigma . \left(\frac{4}{3n}\right)^{1/5}$$

Kernel Density Estimation:

$$f_h(x) = \frac{1}{n\sqrt{2\pi}} \sum_{i=1}^{n} exp\left(-\frac{(x - x_i)^2}{2h^2}\right)$$

Definitions:

- n: Number of data points.
- x_i: Individual data value.
- h: Bandwidth, computed using Scott's Rule
- x: Evaluation point, iterated from min − 3*h* to max + 3*h* in steps of (max − min)/100.
- The bandwidth multiplier in the code may adjust the KDE's scale for optimal display in the Excel plot.

How to Interpret a Violin Plot

A **violin plot** provides a rich visualization of data distribution by combining key elements of a **box plot** and a **kernel density estimate (KDE)**. Here's how to interpret its components:

Shape & Width

- The **violin's width** at any point represents the **probability density** of the data at that value—wider sections indicate higher data concentration.
- **Multiple peaks** suggest **multimodality** (multiple subgroups within the data).
- **Symmetry vs. Skew**:
    - A symmetric violin implies a balanced distribution (e.g., normal distribution).
    - A skewed or asymmetric shape indicates uneven spread (e.g., right/left skew).

Central Tendency & Spread

- **Median:** The central value (50th percentile), splitting the data into equal halves.
- **If mean ≈ median**: The distribution is likely **symmetric** (e.g., normal distribution).
- **If mean > median**: The data is **right-skewed** (tail extends to higher values).
- **If mean < median**: The data is **left-skewed** (tail extends to lower values).

When to Use a Violin Plot vs. a Box Plot

- **Violin plots** excel when you need to:
    - Reveal **subtle distribution patterns** (e.g., bimodality, skew).
    - Compare **density across groups** (e.g., in biology, social sciences).
- **Box plots** are better for:
    - Simple **summary statistics** (median, IQR, outliers).
    - Large datasets where density estimation might be noisy.