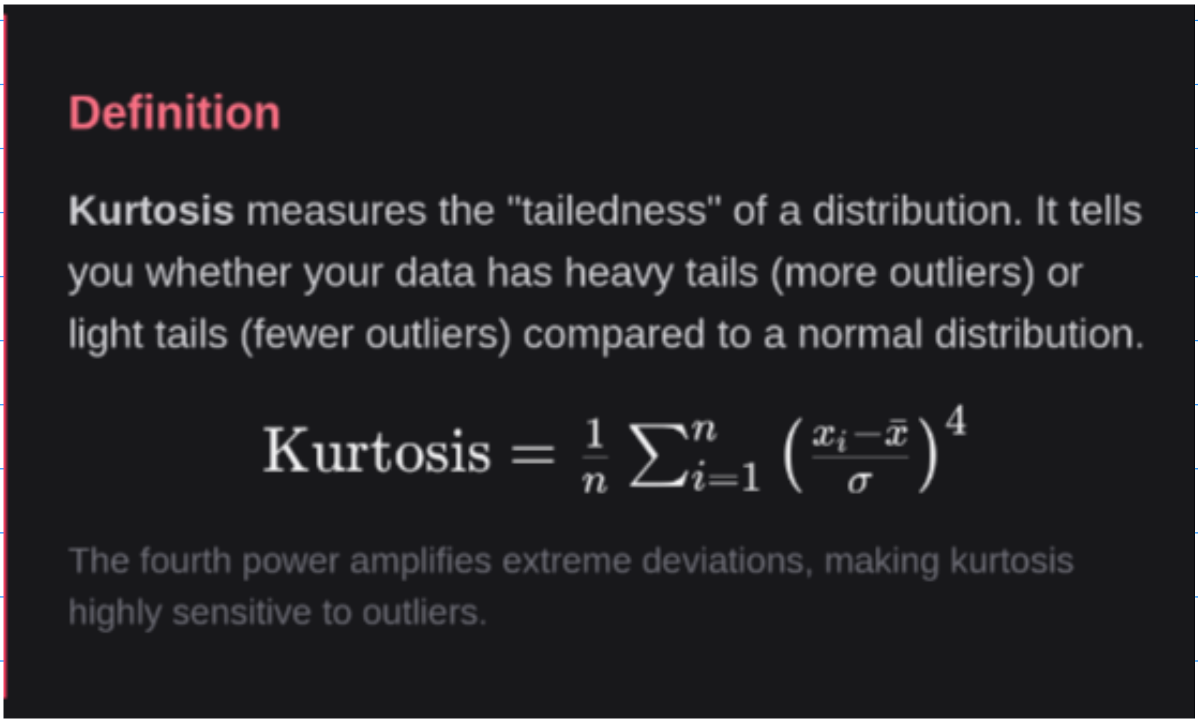
TensorTonic: Statistic - *Skewness Cont'd*

Kurtosis: While Skewness tells about the the asymmetry of our data, kurtosis tells us about the tails. Specifically: How likely are extreme values(outliers) compared to a normal distribution?.



Definition

Kurtosis measures the "tailedness" of a distribution. It tells you whether your data has heavy tails (more outliers) or light tails (fewer outliers) compared to a normal distribution.

$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^4$$

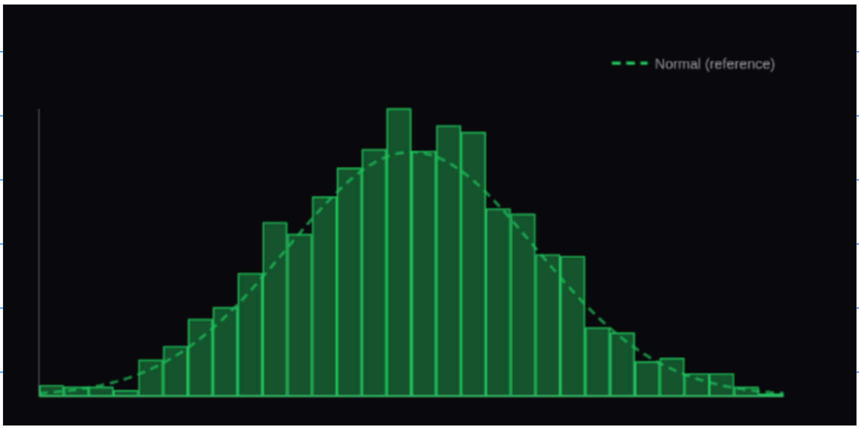
The fourth power amplifies extreme deviations, making kurtosis highly sensitive to outliers.

1. When the value of the calculated Kurtosis is less than 3, it call Platykurtic i.e Flatter peak, thinner tails. Fewer extreme values than normal.

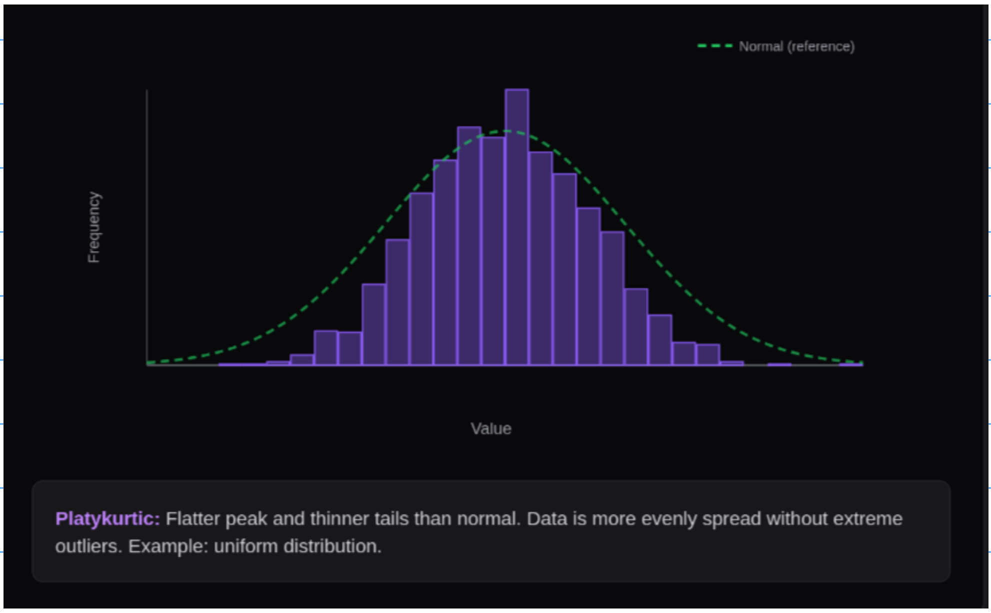
2.When the value is equal to 3, it is called Mesokurtic i.e Normal distribution. The baseline for comparison

3. When the value calculated is greater tham 3, it is called Leptokurtic i.e SHARP Peak, FAT Tails

Mesokurtic:

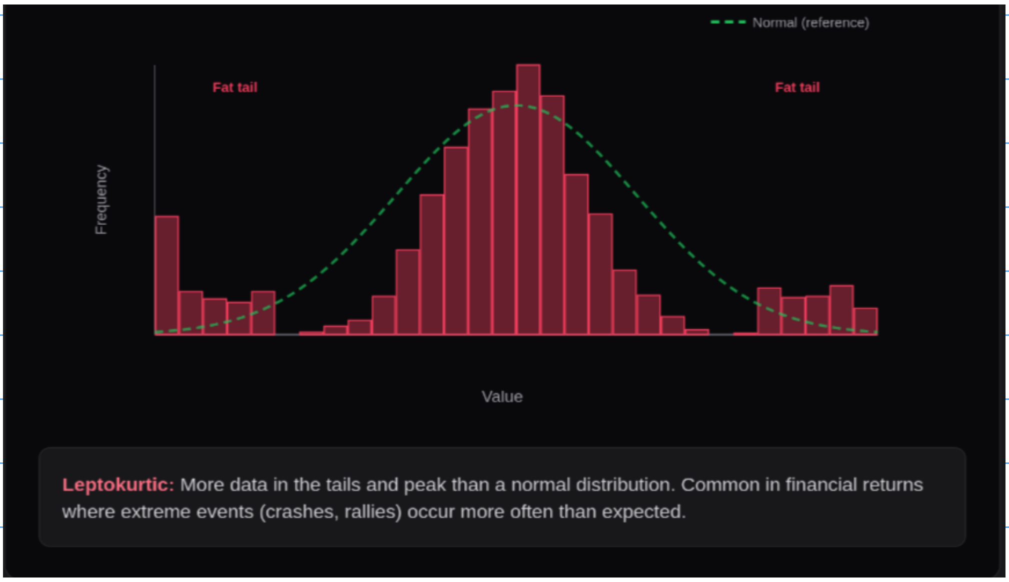


Platykurtic:



Platykurtic: Flatter peak and thinner tails than normal. Data is more evenly spread without extreme outliers. Example: uniform distribution.

Lepokurtic:



Leptokurtic: More data in the tails and peak than a normal distribution. Common in financial returns where extreme events (crashes, rallies) occur more often than expected.

Measure of Spread (Dispersion)

This measures how far, on average, the data points are from the center.

Imagine two companies with the exact same average salary of $100k. In Company A, everyone earns $100k. In Company B, the CEO earns $1M and everyone else earns $10k. The average is the same, but the spread tells the real story.So how do we measure this? We need a way to calculate how far, on average, the data points are from the center.

Range: The difference between the maximum point and minimum point in the dataset. Its is the Simplest measure of gap/spread.

$$Range = x_{max} - x_{min}$$

Weakness: It is extremely sensitive to outliers.

Interquartile Range (IQR): This is the range of the middle i.e 50% of the data. It is given as

$$IQR = Q_3 - Q_1$$

where the Q_3 represent the third quartile and the Q_1 represent the first quartile.

TIP: Outlier Detection: A common rule is to flag any data point as an outlier if it falls below $IQR = Q_1 - 1.5 \times IQR$ or above $IQR = Q_3 + 1.5 \times IQR$ .This is how box plots draw their "whiskers."

Variance: Variance looks at every single data point and calculates how far it sits from the average. We square those distances (so the negative numbers don't cancel out the positive ones) and average them. It's one of the most used in statistic and machine learning.

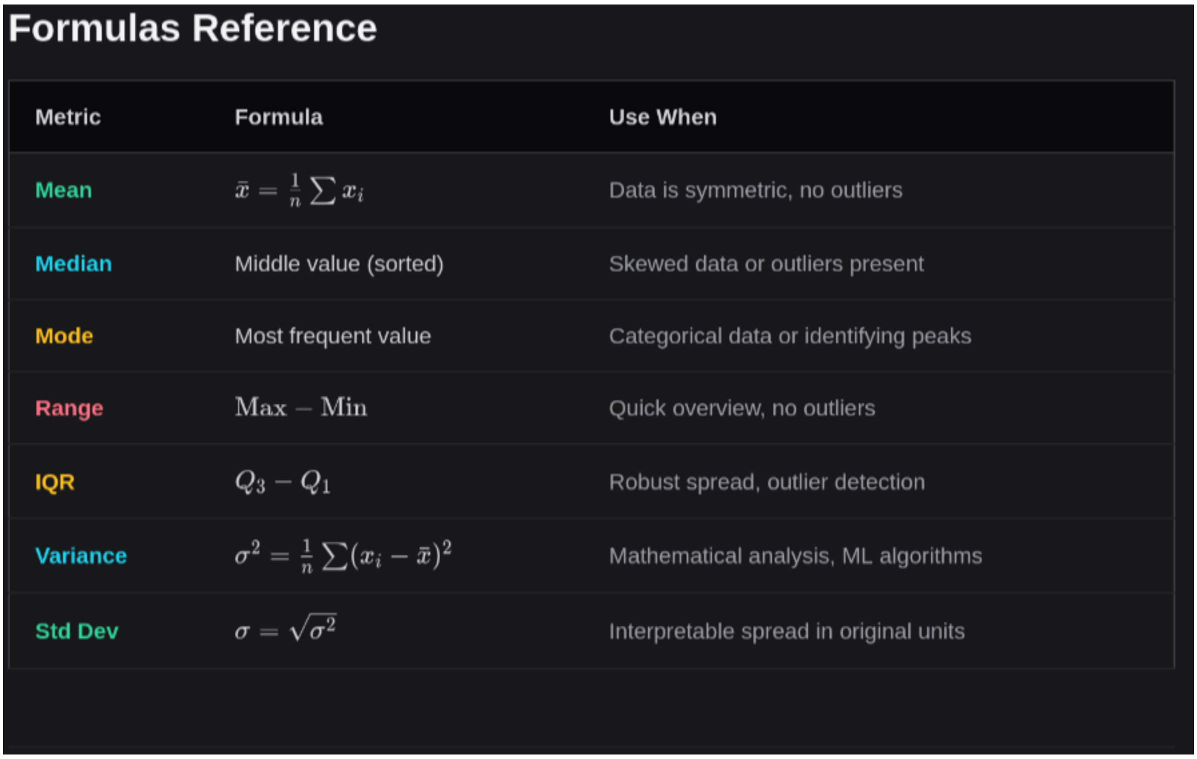
$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Where:

N: The Population or Sample Size.
x_i: The data point/ instance

Standard Deviation: This is just the square root of the variance. Why do we do this extra step? Because variance is measured in "squared units" (like "squared degrees"), which makes no sense to humans. Standard deviation puts the answer back into the original units (like "degrees").

More on outliers: The Range is a bit of a drama queen. If you have a room of middle-class people and Elon Musk walks in, your "Range" of wealth suddenly explodes, even though nobody else's bank account changed. To fix this,we use the Interquartile Range (IQR), which ignores the top 25% and bottom 25% and only looks at the "middle" 50% of the data.



Formulas Reference

Metric	Formula	Use When
Mean	$\bar{x} = \frac{1}{n} \sum x_i$	Data is symmetric, no outliers
Median	Middle value (sorted)	Skewed data or outliers present
Mode	Most frequent value	Categorical data or identifying peaks
Range	$\text{Max} - \text{Min}$	Quick overview, no outliers
IQR	$Q_3 - Q_1$	Robust spread, outlier detection
Variance	$\sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$	Mathematical analysis, ML algorithms
Std Dev	$\sigma = \sqrt{\sigma^2}$	Interpretable spread in original units

Population VS Sample