Data aggregations and descriptive statistics

Measures of central tendency ⌄

A single value that seeks to describe a set of data by identifying the typical value.

Mean

The value that identifies the centre by calculating the **arithmetic average** of all the data points.

Commonly used when the data do not contain outliers since the mean is very sensitive to outliers.

`=AVERAGE(value1, [value2, ...])` </>

$$\text{Mean} = \frac{\text{Sum of all data points}}{\text{Number of data points}}$$

$$= \frac{1+3+5+8+10+13+16}{7} = \boxed{8}$$

Median

The value that is exactly **in the middle** of a set of values after we have ordered the values from smallest to largest.

Commonly used when the data contain a lot of outliers.

`=MEDIAN(value1, [value2, ...])` </>

{ 1, 3, 5, 8, 10, 13, 16 }   Median = 8

Mode

The value that appears **most frequently** in the dataset.

Commonly used when the data are categorical. That is, the data contain a fixed number of groups.

`=MODE(value1, [value2, ...])` </>

{ 5, 1, 5, 8, 5, 1, 5 }   Mode = 5
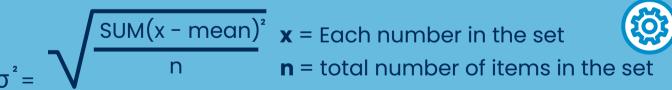
Measures of spread ⌄

Describe how far the values of a dataset lie from each other and from the mean, median or mode.
A measure of spread gives us an idea of how well the mean, for example, describes the data.

Standard deviation

Measures the amount of **variation** that exists in the dataset by calculating the difference between **all data points and the mean**.

`=STDEV(value1, [value2, ...])` </>

$$\sigma^2 = \sqrt{\frac{\text{SUM}(x - \text{mean})^2}{n}}$$

x = Each number in the set
n = total number of items in the set

Variance

Measures the amount of **variation** that exists in the dataset by calculating the difference between **all data points and the mean**.

`=VAR(value1, [value2, ...])` </>

$$\sigma^2 = \frac{\text{SUM}(x - \text{mean})^2}{n}$$

x = Each number in the set
n = total number of items in the set

Interquartile range

The value that measures the **spread of the middle half** of the data.

Assesses the variability of the middle 50% of the data, where we assume the majority of the values lie.

`=QUARTILE(data, 3) - QUARTILE(data, 1)` </>

Q1   Median   Q2
{ 1, 3, 5, 6, 8, 10, 13, 16, 17, 19, 22 }   17 – 5 = 12

Range

The value that measures the spread by considering the **difference between the smallest and largest** values.

`=MAX(value1, [value2, ...])`
`  - MIN(value1, [value2, ...])` </>

{ 1, 3, 5, 8, 10, 13, 16 }   16 – 1 = 15

Quartiles

The quartiles segment the data distribution into **four equal quarters**, with the median being in the middle of these quartiles.

`=QUARTILE(data, quartile_number)` </>

Q1   Median   Q2

{ 1,   3,   5,   6,   8,   10,   13,   16,   17,   19,   22 }