Summer of Code
Artificial Intelligence
(Machine Learning & Deep Learning)

Instructor
Wajahat Ullah
- *Research Assistant* (DIP Lab)
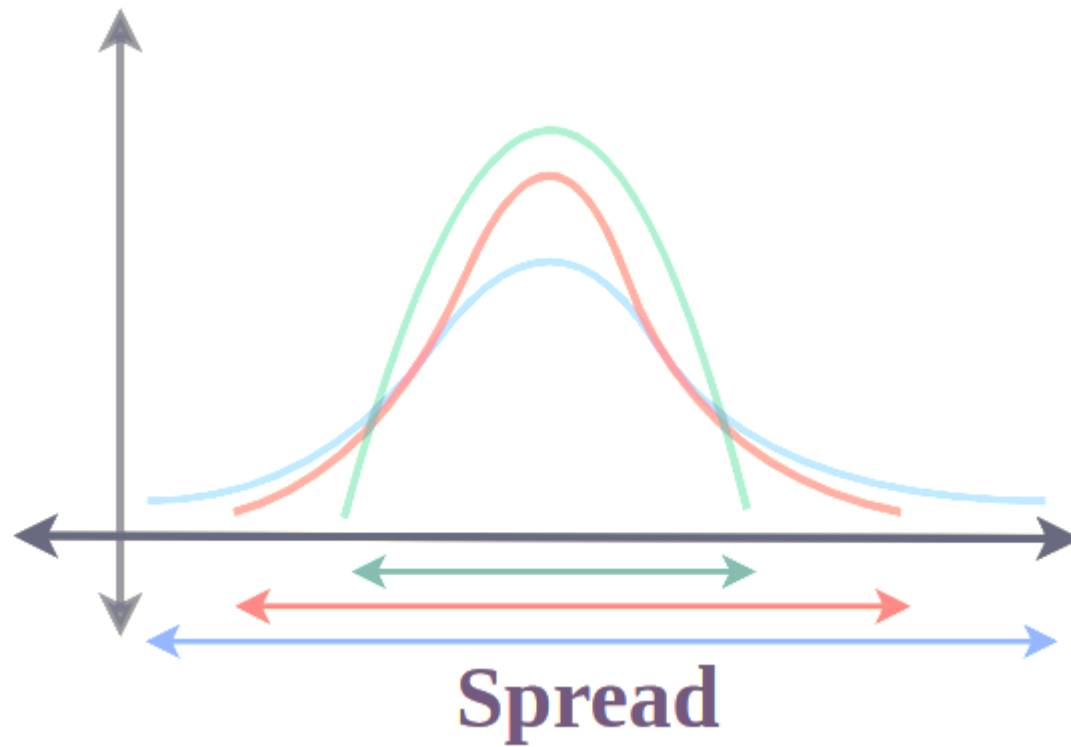
Duration
03 Months
(September – November)

Day 04 – Descriptive Statistics
(Measure of Dispersion and Position)

Objectives:

❖ What is Dispersion or Variability

❖ Measures of Dispersion

❖ Measures of Position

Dispersion

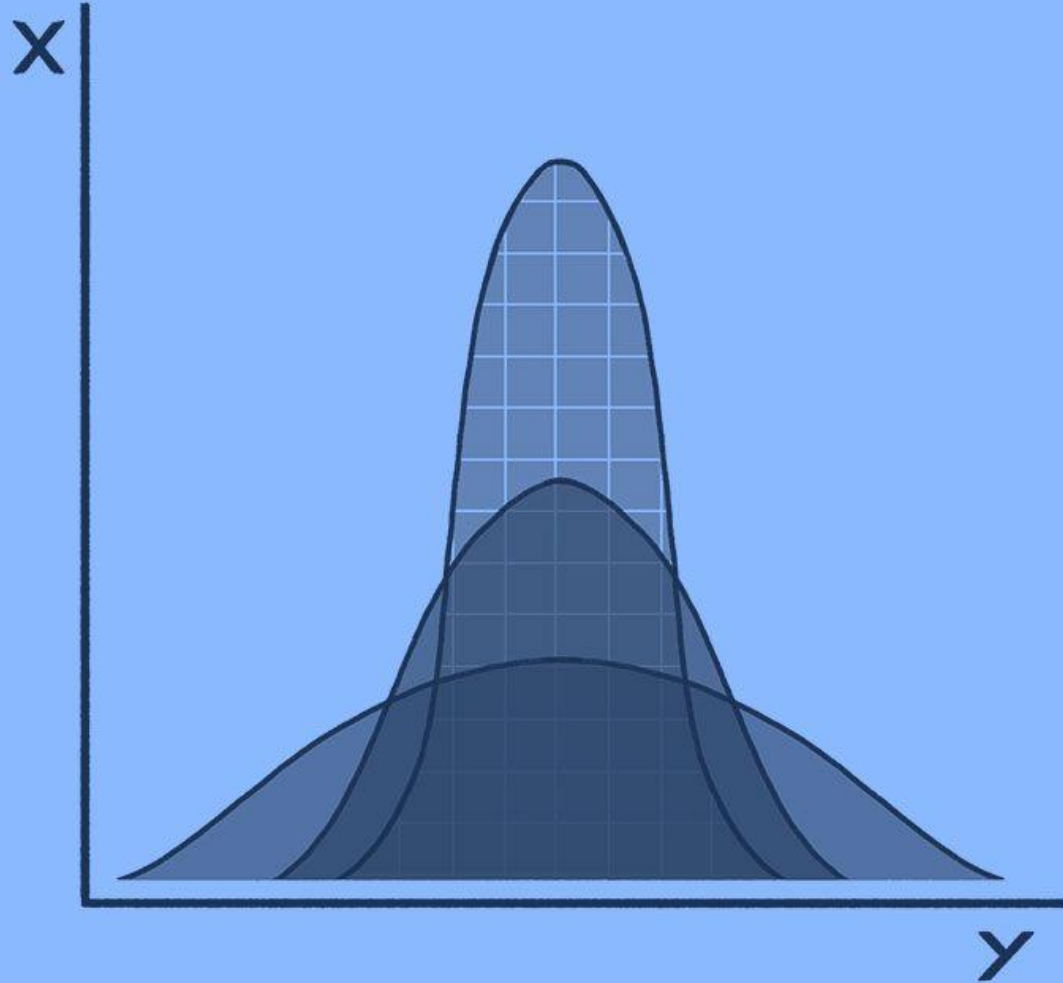Dispersion is the spread or variability of the data points from the mean value

Spread

FIVE MEASURES OF DISPERSION

1 — RANGE
Difference between the maximum and minimum values

2 — VARIANCE
Average of the squared differences from the mean

3 — STANDARD DEVIATION
Square root of the variance

4 — MEAN ABSOLUTE DEVIATION
Average of the positive difference from the mean

5 — INTERQUARTILE RANGE
Focuses on the middle 50% of the dataset

Range

The difference between the largest value
and the smallest value is the range.

1   2   3   4   5

Smallest Number   **Largest Number**

Variance

['ver-ē-ən(t)s]

A measurement of how far each number in a data set is from the mean (average), and thus from every other number in the set.
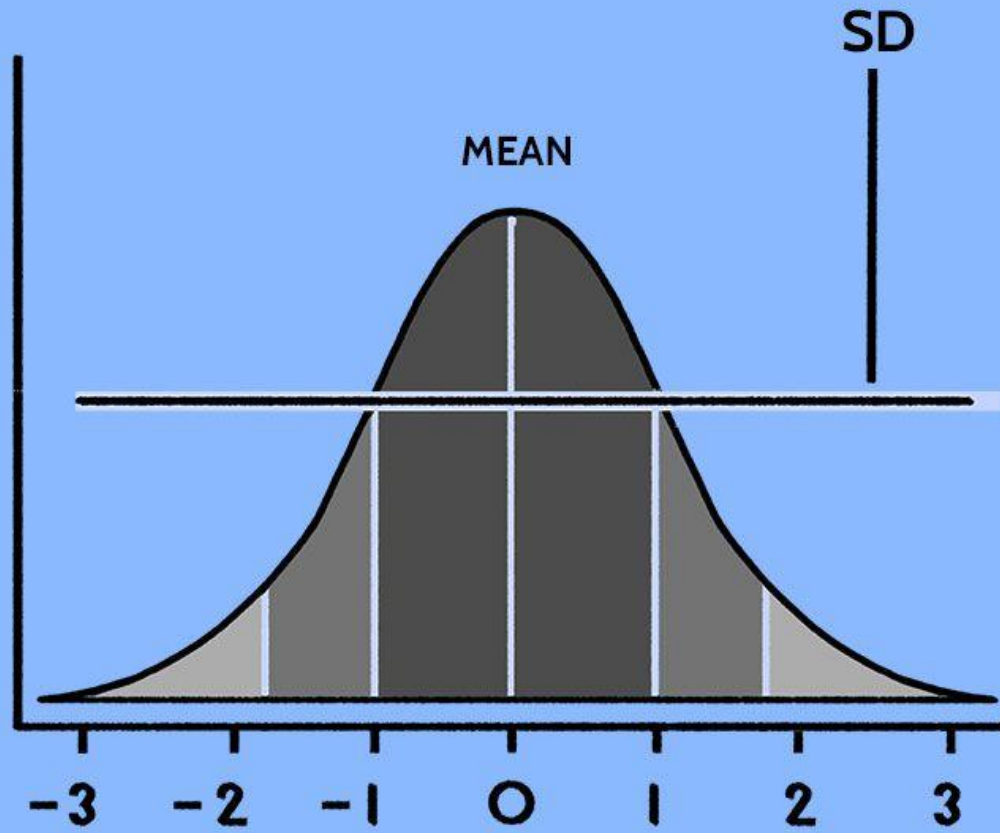
Calculating Variance

$$\text{Variance} = \sigma^2 = \frac{\sum(xi - \bar{x})^2}{N}$$

n = Number of Observations

X_i = Value of the one Observation

$\bar{X}$ = Mean Value of Observations

Standard Deviation

['stan-dərd dē-vē-'ā-shən]

A statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.
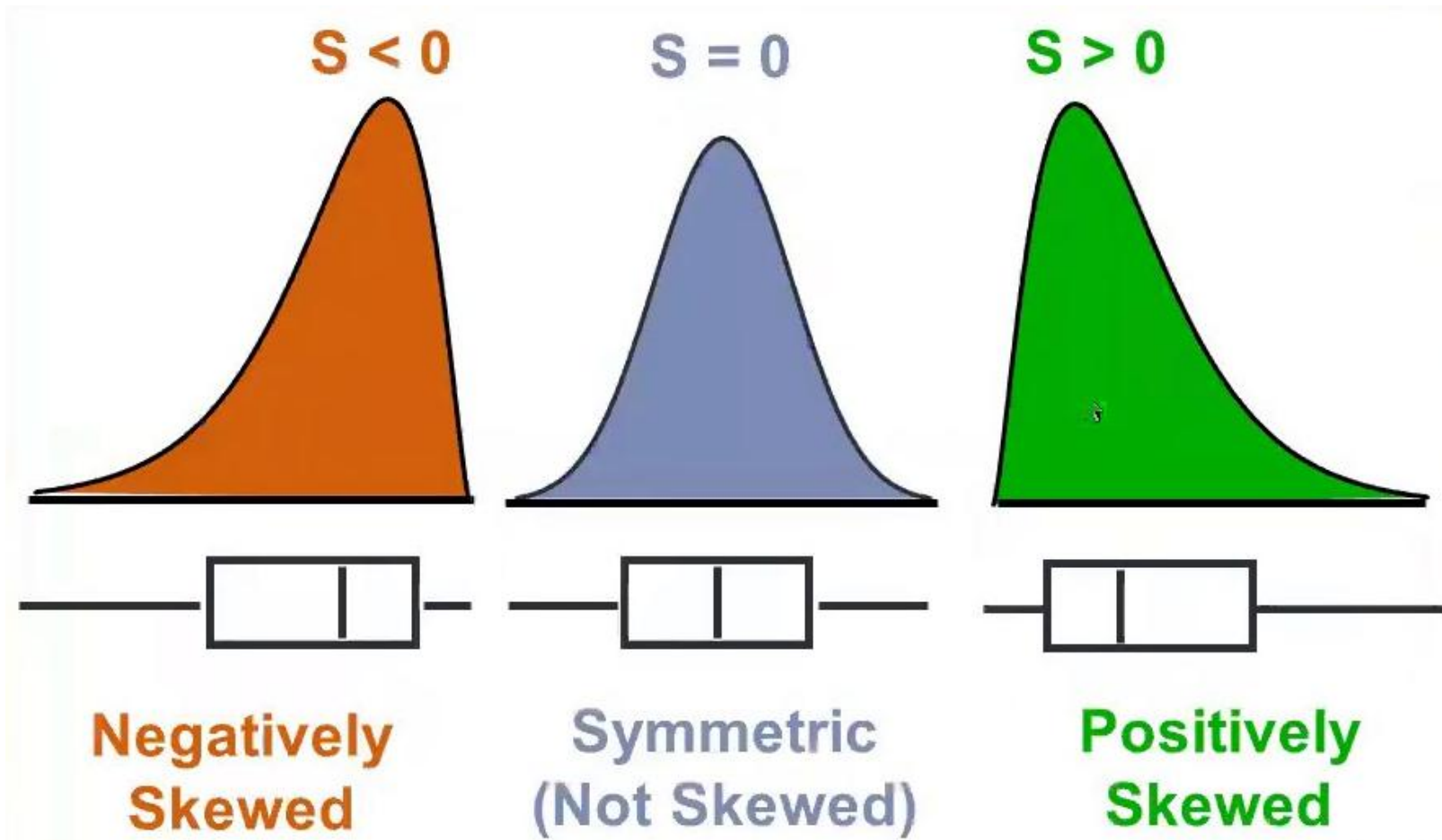
Calculating Standard Deviation

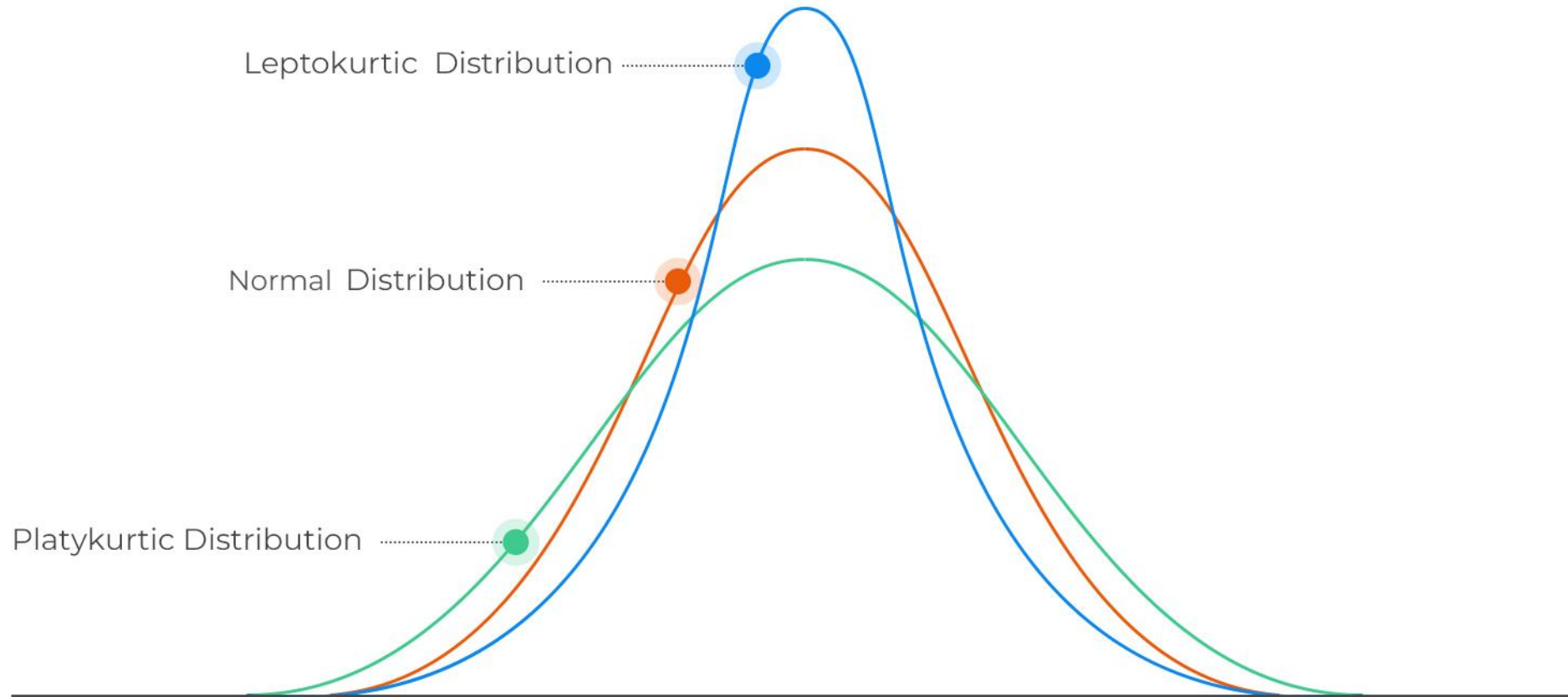Standard Deviation = $\sigma = \sqrt{Variance}$

Skewness

Skewness is a measure of the asymmetry of the data distribution about its mean. The skewness value can be positive, zero, negative
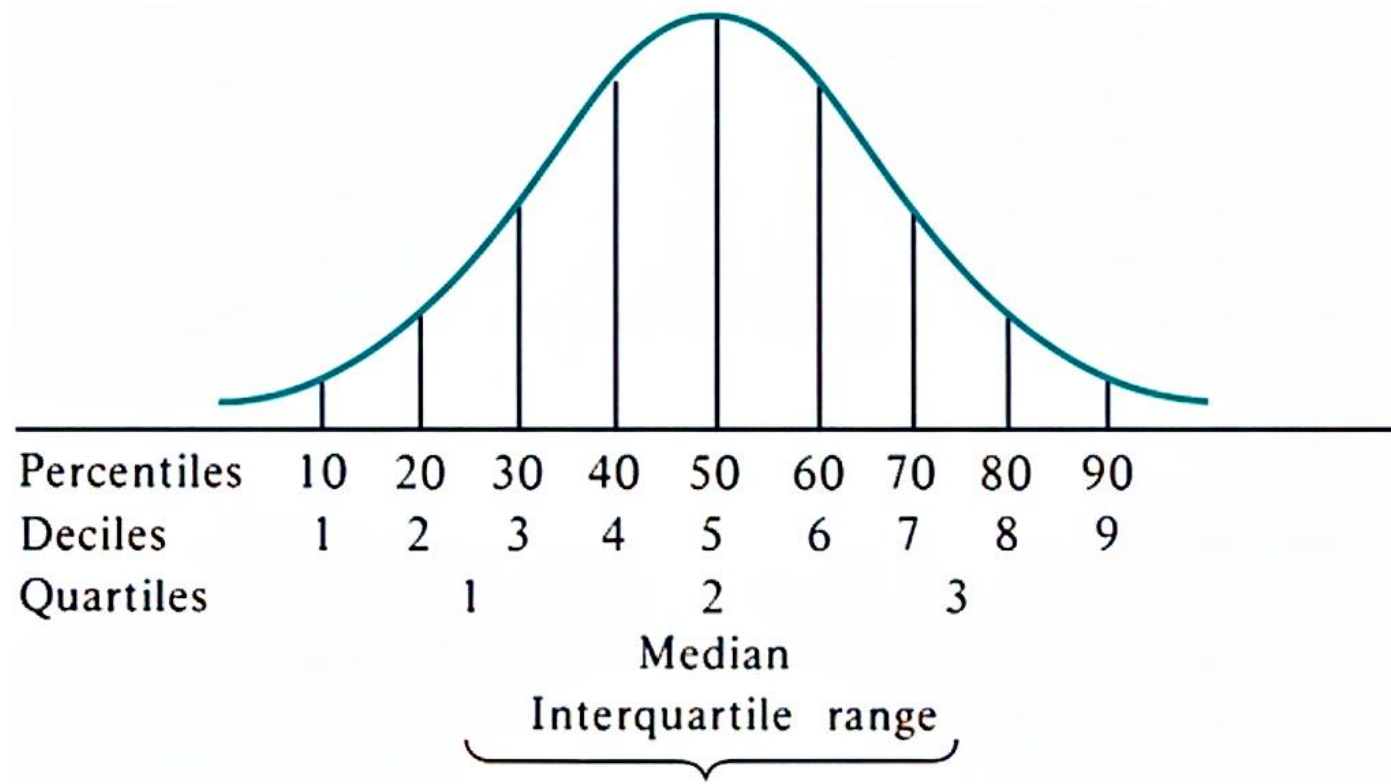
Kurtosis

Kurtosis describes the degree to which scores cluster in the tails or the peak of a data distribution.
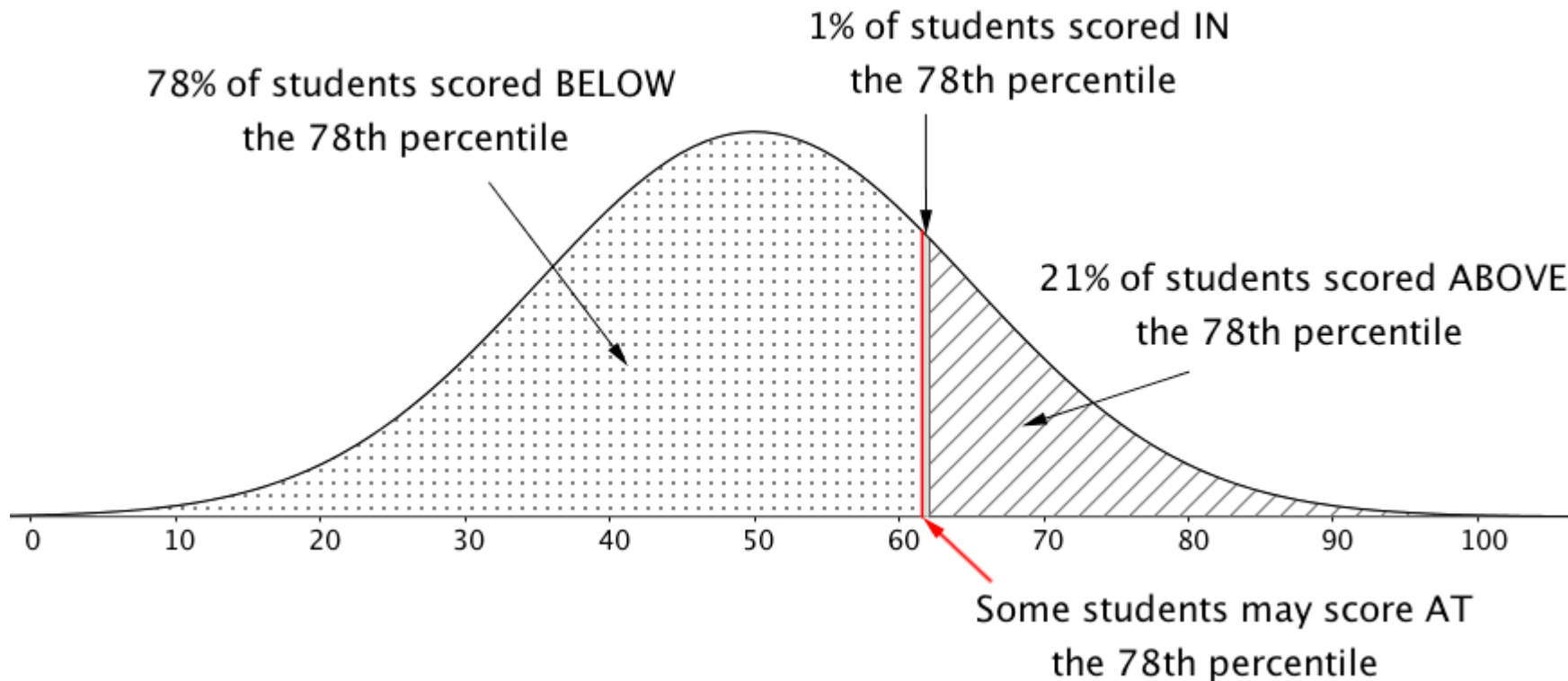
Measures of Position

Numerical measures that describe the standing or location of an observation relative to the rest of the data
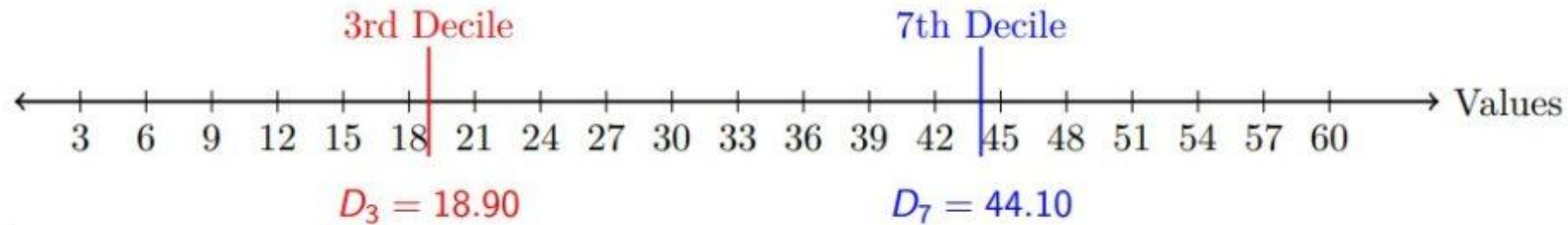
Percentile

Percentiles indicate the value below which a given percentage of observations in a dataset fall. For example, the 25th percentile (P_{25}) is the value below which 25% of the data points lie.
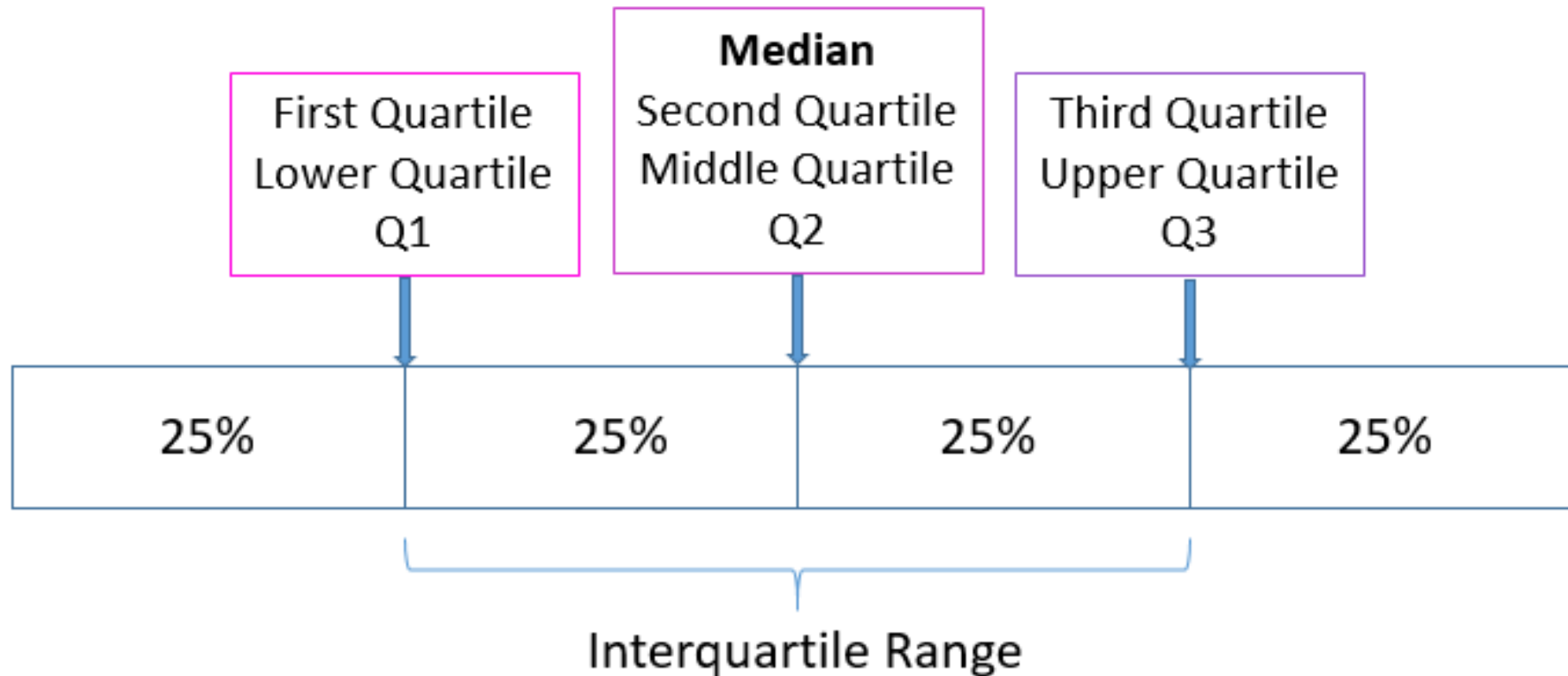
Decile

Deciles divide ranked data into 10 equal parts. Each part is called a decile.

3rd Decile

$D_3 = 18.90$
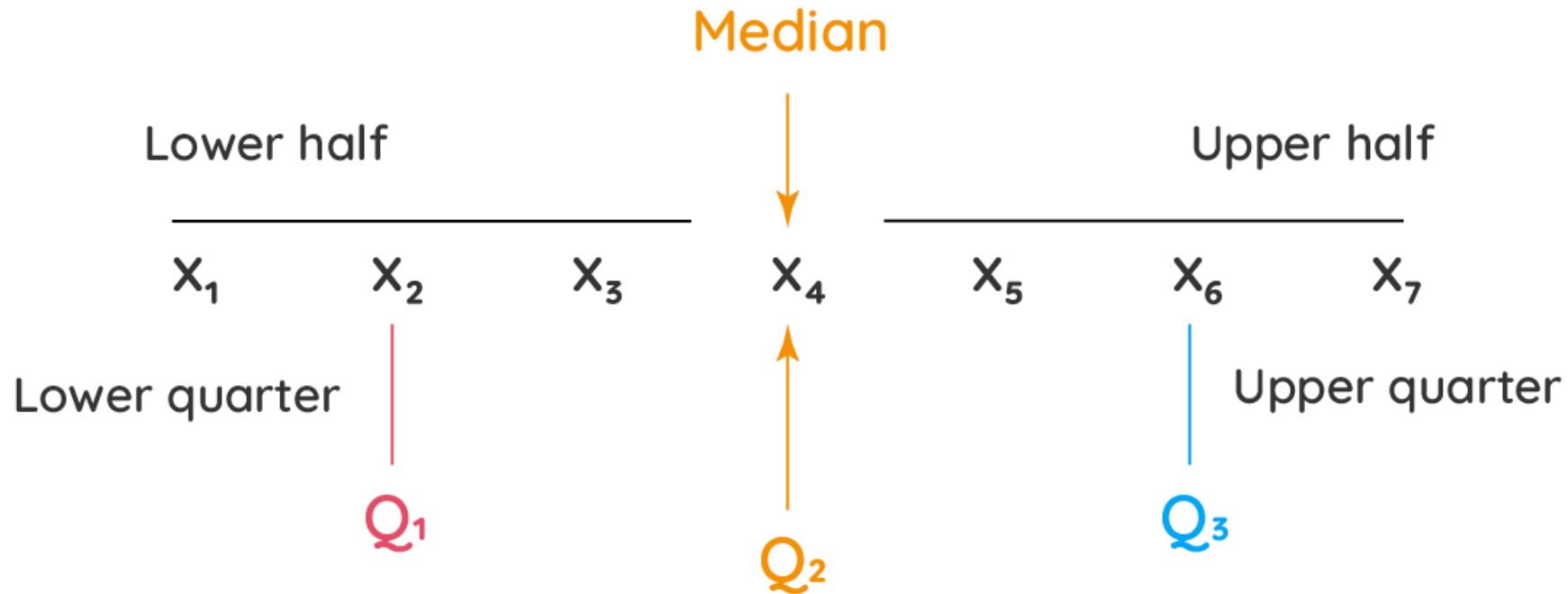
7th Decile

$D_7 = 44.10$

Quartile

Quartiles are values that divide a sorted dataset into four equal parts or quarters, each representing 25% of the data.

Interquartile Range

It is a measure of dispersion of data that represents the spread of the middle 50% of a dataset.

Median

Lower half

Upper half

X_1 X_2 X_3 X_4 X_5 X_6 X_7

Lower quarter

Upper quarter

Q_1

Q_2

Q_3

Interquartile Range: $Q_3 - Q_1$

The Normal Distribution

Also know as Gaussian distribution or Bell Curve, is a continuous probability distribution that is symmetrical and bell-shaped, with data points clustering around the mean and decreasing symmetrically on either side.
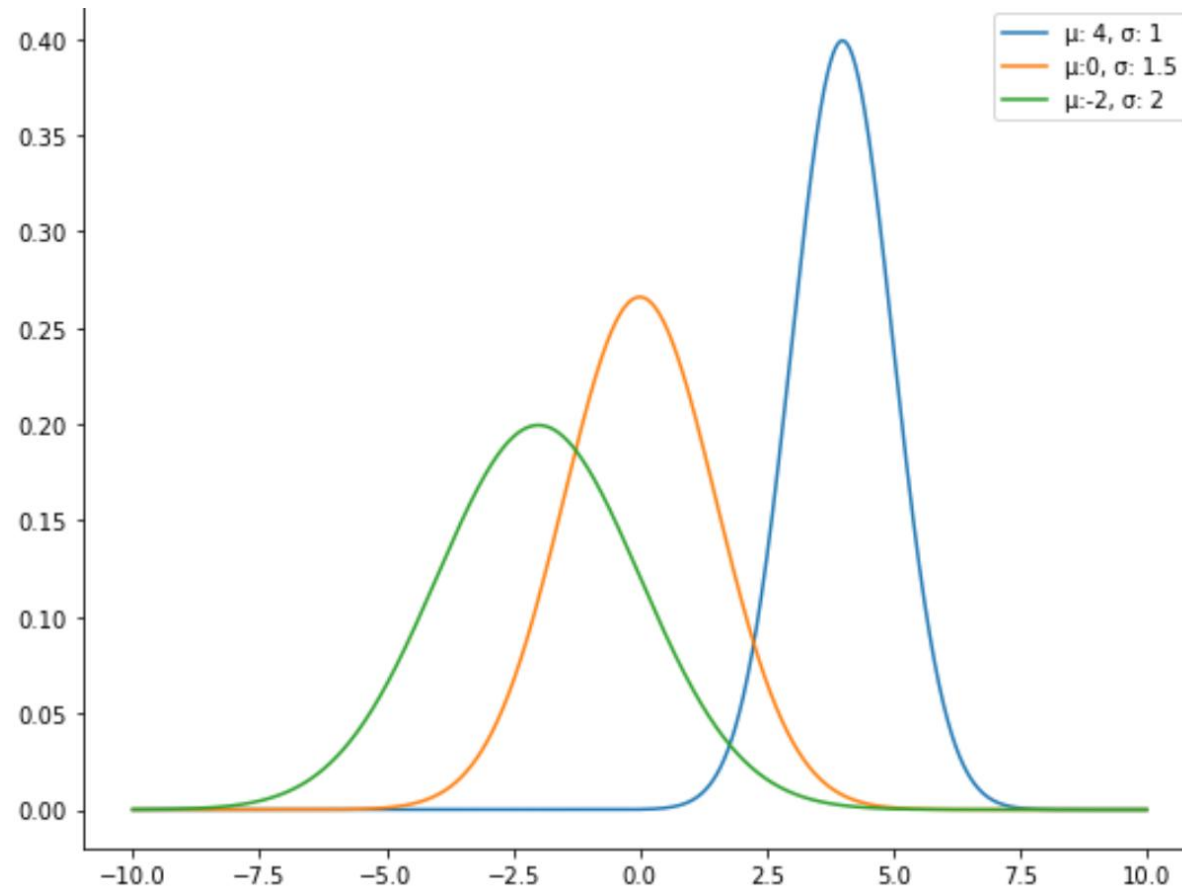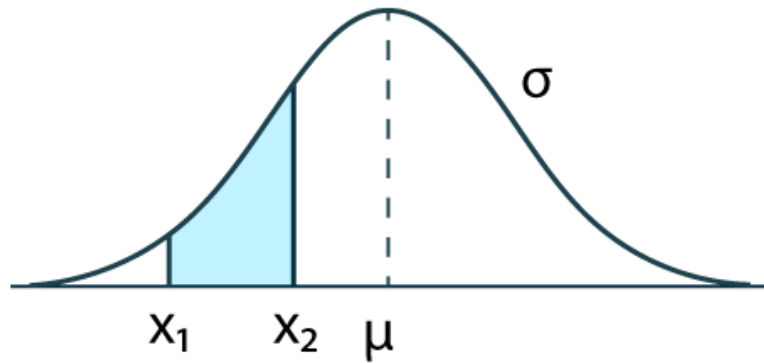
The 68-95-99.7% Rule

Approximately 68% of data falls within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations
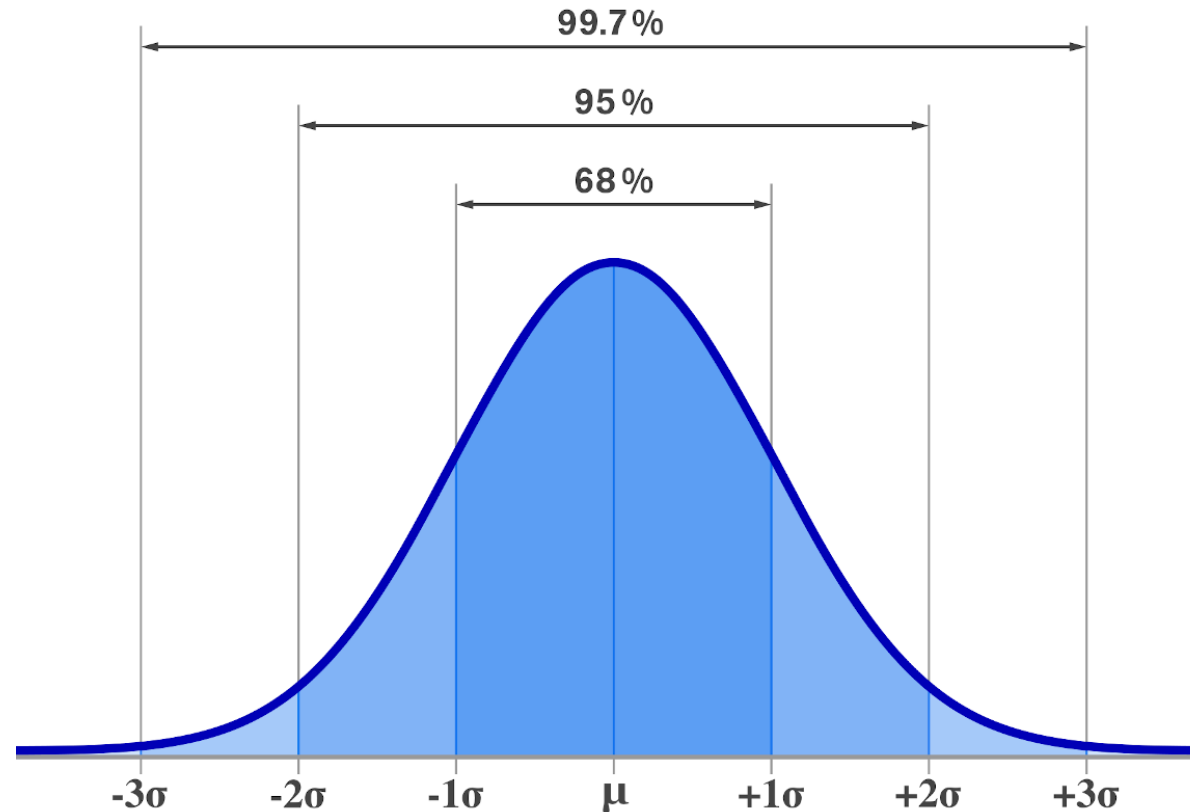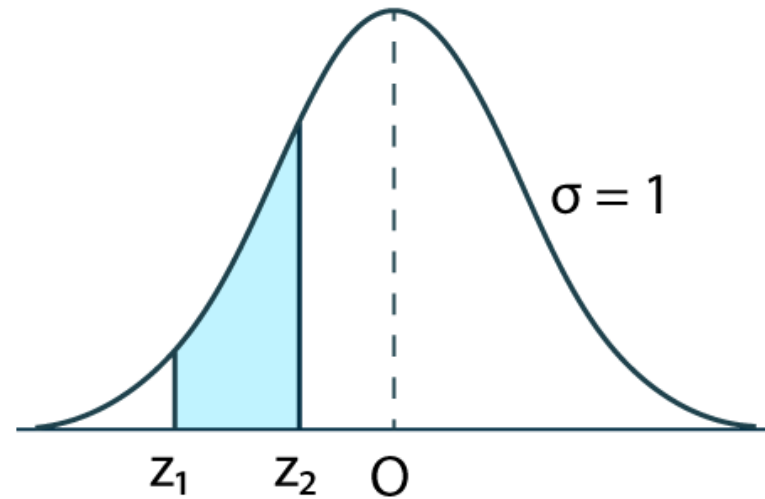
Standard Normal Distribution

It is a specific normal distribution with a mean of 0 and a standard deviation of 1, often used to compare normal distributions

Standard Distribution

Standard Normal Distribution
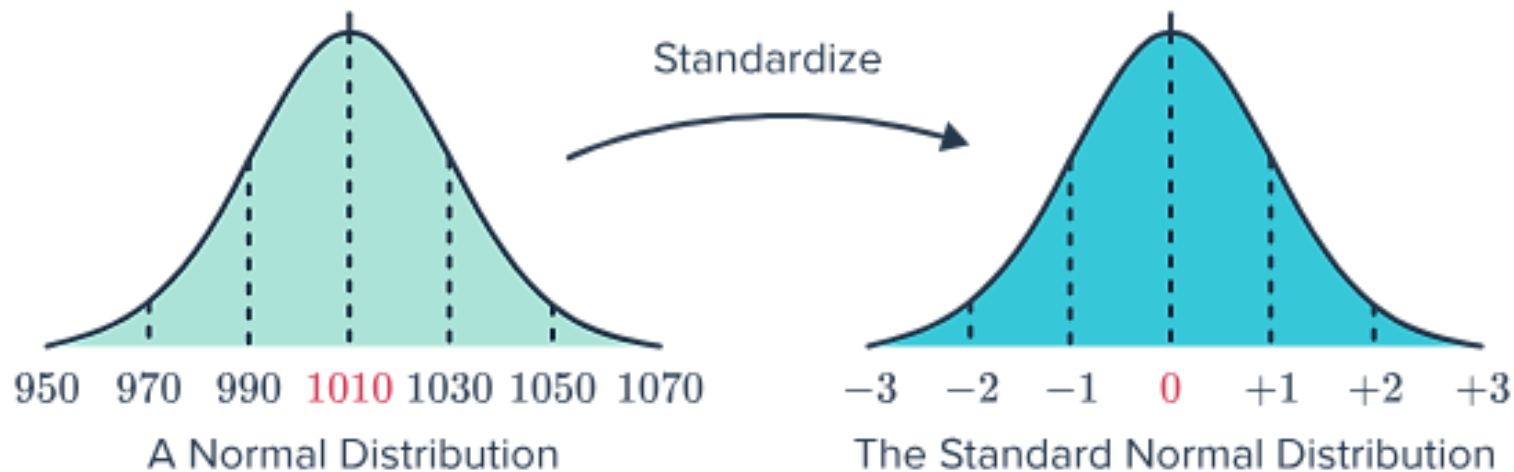
Standardization (Z-Scores)

The z-score indicates how many standard deviations a measurement is from the mean
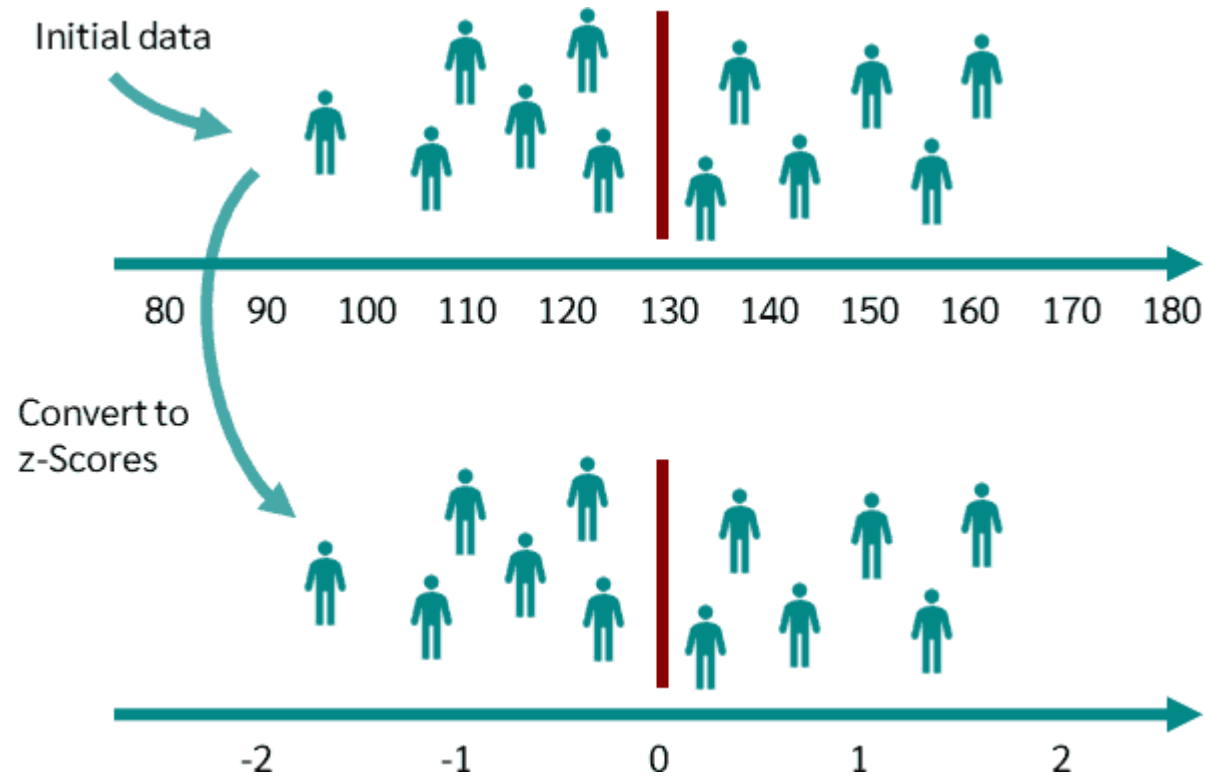
$$Z = \frac{X - \mu}{\sigma}$$

Standardize

950   970   990   1010 1030 1050 1070   −3   −2   −1   0   +1   +2   +3

A Normal Distribution   The Standard Normal Distribution

Standard Normal Distribution
(Z-Scores)

The z-score indicates how many standard deviations a measurement is from the mean

Standard Normal Distribution (Z-Scores)

The z-score indicates how many standard deviations a measurement is from the mean

Thank You