UNIT 04

Measures of Dispersion/Variability

MATH -2205

Measures of Dispersion or Variability

Measures of dispersion, also known as measures of variability, are statistical tools used to describe the spread or distribution of a set of data. These measures indicate how much the data points differ from the central tendency (mean, median, mode) and from each other.

Studying measures of dispersion is essential for gaining a deeper understanding of data. They provide critical information about data variability, enhance the accuracy of statistical analyses, support decision-making processes, and improve the reliability of predictive models. Without these measures, we would have an incomplete picture of the data, potentially leading to incorrect conclusions and poor decisions.

Common Measures of Dispersion or Variability

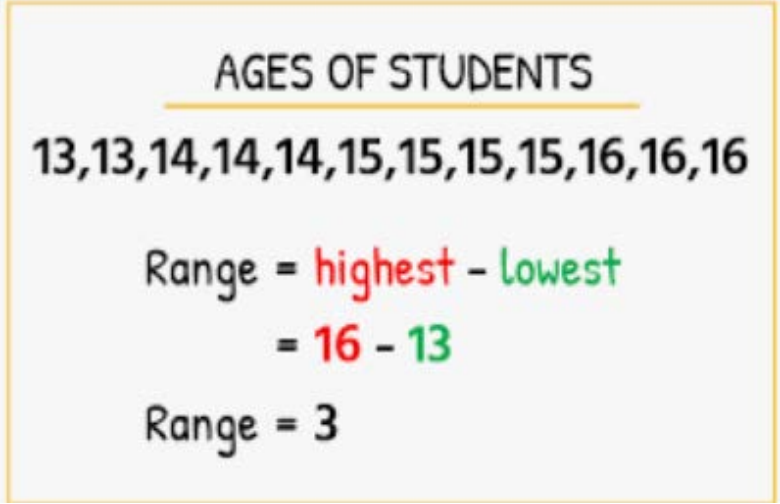
- Range
- Interquartile Range (IQR)
- Mean Absolute Deviation
- Variance
- Standard Deviation
- Coefficient of Variation (CV)

Range

The range is a measure of dispersion in statistics that represents the difference between the highest and lowest values in a dataset.

It provides a quick sense of the spread of the data.

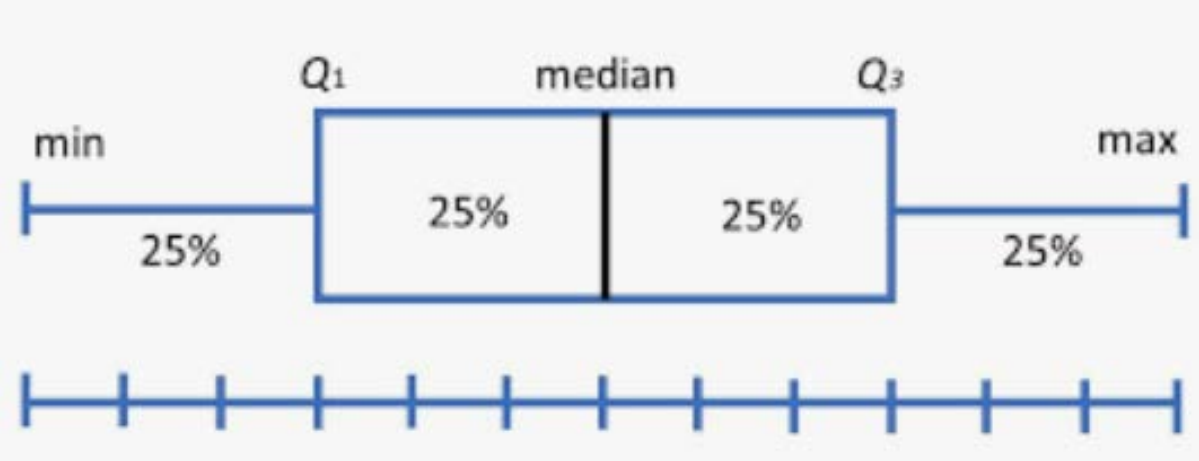
Range= Maximum Value−Minimum Value



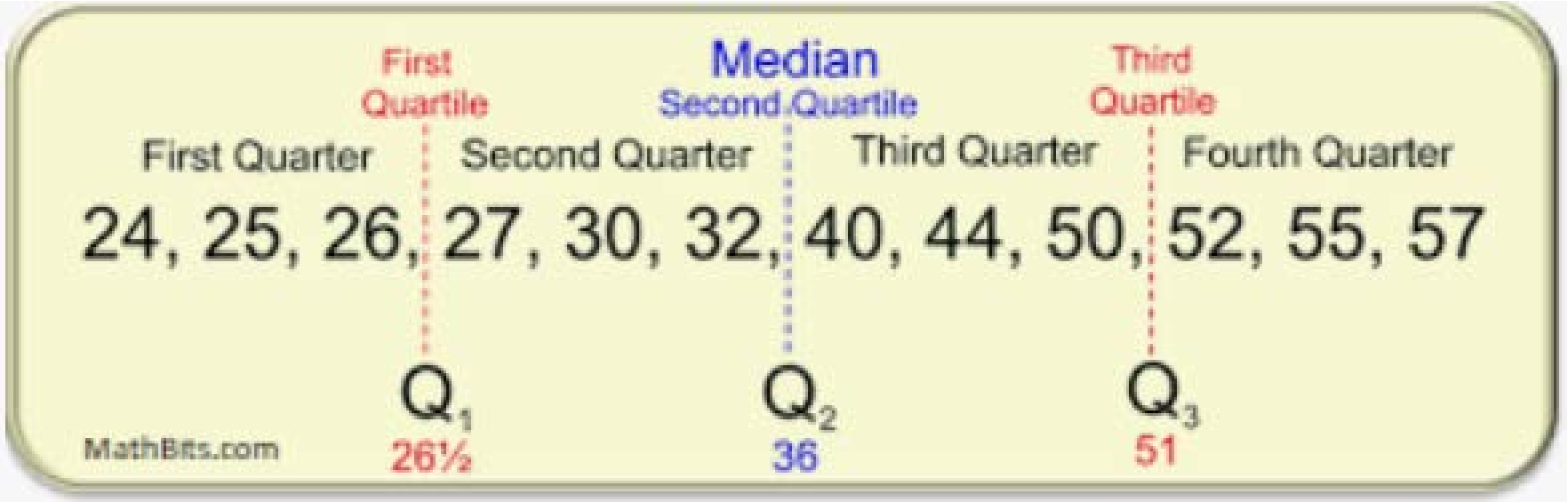
Interquartile Range (IQR)

Quartiles:

Quartiles are statistical measures that divide a dataset into four equal parts, each containing 25% of the data points. They provide insights into the distribution and spread of the data, helping to identify the central tendency and variability.



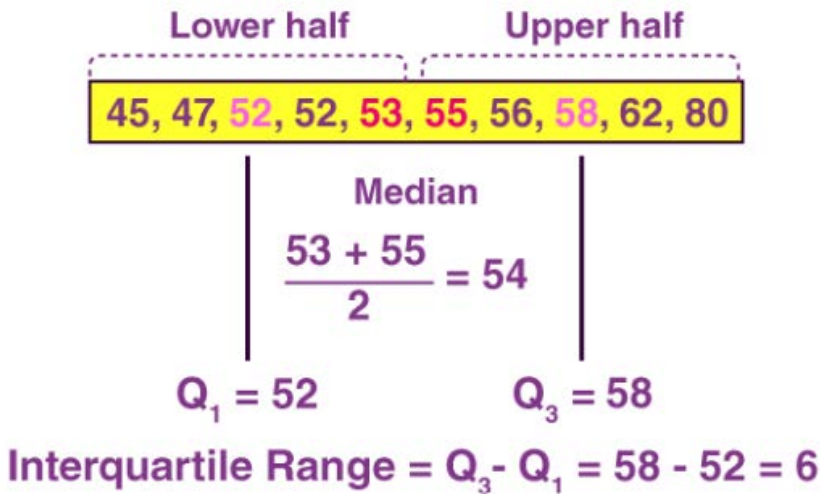
Example


First Quartile
Median
Second Quartile
Third Quartile
First Quarter
Second Quarter
Third Quarter
Fourth Quarter
24, 25, 26, 27, 30, 32, 40, 44, 50, 52, 55, 57
Q_1
Q_2
Q_3
MathBits.com
26½
36
51

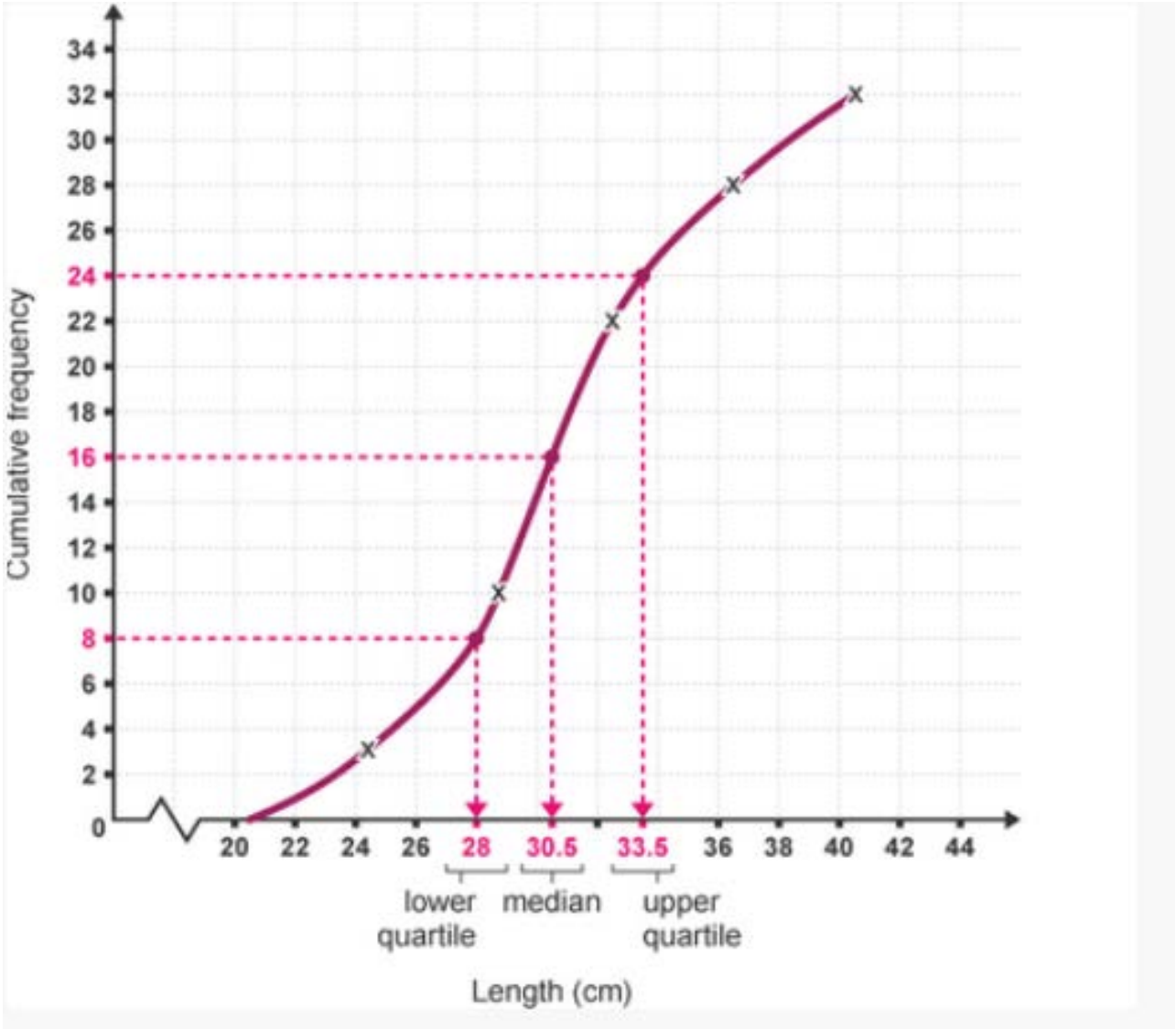
Interquartile Range (IQR)

The Interquartile Range (IQR) is a measure of statistical dispersion, which is the spread of the middle 50% of the data points. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$$\text{IQR} = Q_3 - Q_1$$



Calculating IQR using cumulative frequency graph



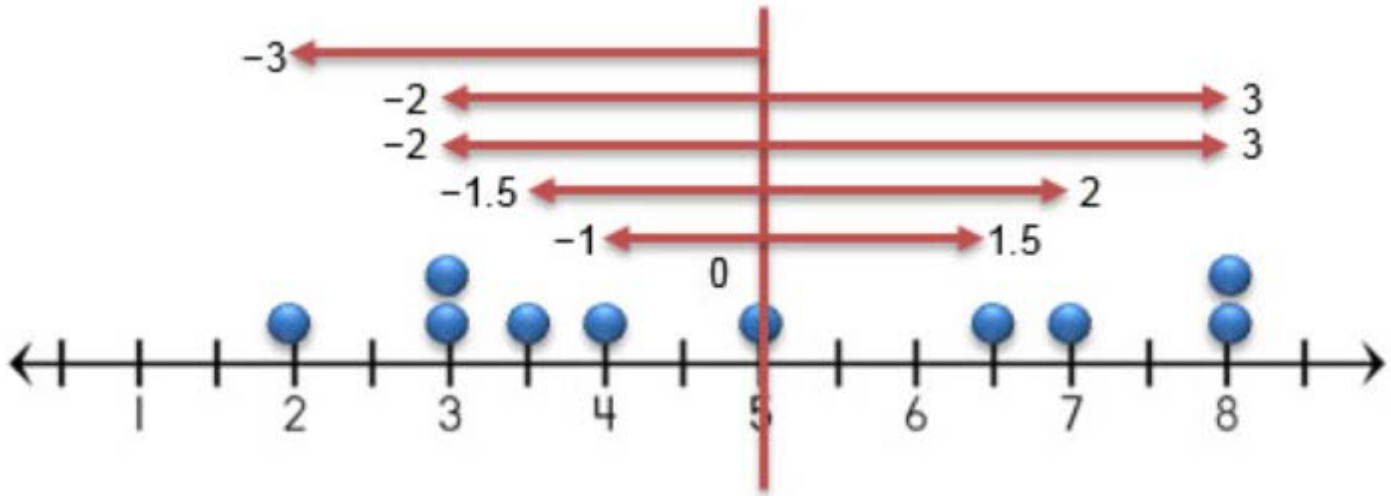
Mean Absolute Deviation

Mean deviation, also known as the average absolute deviation, is a measure of dispersion that indicates the average distance between each data point and the central value (mean or median) of the dataset. It provides a summary of the variability within a dataset by considering the absolute differences between data points and a central measure.

$$\text{Mean deviation} = \frac{\sum|x-\bar{x}|}{n}$$

$$\frac{1}{n}\sum_{i=1}^{n}|x_i-m| = \frac{|x_1-m|+|x_2-m|+\cdots+|x_n-m|}{n}$$

Where m is the mean



Mean = 5

x	$x - \mu$	$\lvert x - \mu \rvert$
6	-3	3
7	-2	2
10	1	1
12	3	3
13	4	4
4	-5	5
8	-1	1
12	3	3

Variance

Variance is a measure of how much the values in a dataset differ from the mean of the dataset. It quantifies the degree of dispersion or spread in the data. A high variance indicates that the data points are spread out widely from the mean, while a low variance indicates that they are clustered closely around the mean.

For a population:

$$\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$$

For a sample:

$$s^2 = \frac{\sum(x_i-\overline{x})^2}{n-1}$$

Where:

- σ^2 = population variance
- s^2 = sample variance
- x_i = each individual data point
- μ = population mean
- $\overline{x}$ = sample mean
- N = number of data points in the population
- n = number of data points in the sample

Calculation Steps:

1. **Calculate the Mean (Average)**:
   - For a sample: $\overline{x} = \frac{\sum x_i}{n}$
   - For a population: $\mu = \frac{\sum x_i}{N}$
2. **Subtract the Mean from Each Data Point and Square the Result**:
   - $(x_i - \overline{x})^2$ for each data point in a sample
   - $(x_i - \mu)^2$ for each data point in a population
3. **Sum All Squared Differences**:
   - $\sum(x_i - \overline{x})^2$ for a sample
   - $\sum(x_i - \mu)^2$ for a population

4. **Divide by the Number of Data Points:**

   - For a population: Divide by N
   - For a sample: Divide by $n - 1$ (this is called Bessel's correction, which corrects the bias in the estimation of the population variance)

Example

Consider the dataset: 4, 8, 6, 5, 3

Step 1: Calculate the Mean:

$$\overline{x} = \frac{4+8+6+5+3}{5} = \frac{26}{5} = 5.2$$

Step 2: Calculate Each Squared Deviation from the Mean:

$$(4 - 5.2)^2 = 1.44$$
$$(8 - 5.2)^2 = 7.84$$
$$(6 - 5.2)^2 = 0.64$$
$$(5 - 5.2)^2 = 0.04$$
$$(3 - 5.2)^2 = 4.84$$

Step 3: Sum the Squared Deviations:

$$1.44 + 7.84 + 0.64 + 0.04 + 4.84 = 14.8$$

Step 4: Divide by the Number of Data Points (for a sample):

$$s^2 = \frac{14.8}{5-1} = \frac{14.8}{4} = 3.7$$

Standard Deviation

Standard deviation is the square root of variance. Live variance it also measures the amount of variation or dispersion in a set of values. It indicates how much the individual data points deviate from the mean (average) of the dataset. A low standard deviation means the data points are close to the mean, while a high standard deviation indicates that the data points are spread out over a wide range.

For a population:

$$\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{N}}$$

For a sample:

$$s = \sqrt{\frac{\sum(x_i-\overline{x})^2}{n-1}}$$

- σ = population standard deviation
- s = sample standard deviation

Coefficient of Variation (CV)

The Coefficient of Variation (CV) is a statistical measure of the relative dispersion of data points in a dataset. It is expressed as a percentage and is used to compare the degree of variation between different datasets, even if the datasets have different units or means. The CV is particularly useful for comparing the variability of data that have different scales or different units of measurement..

- For a population:

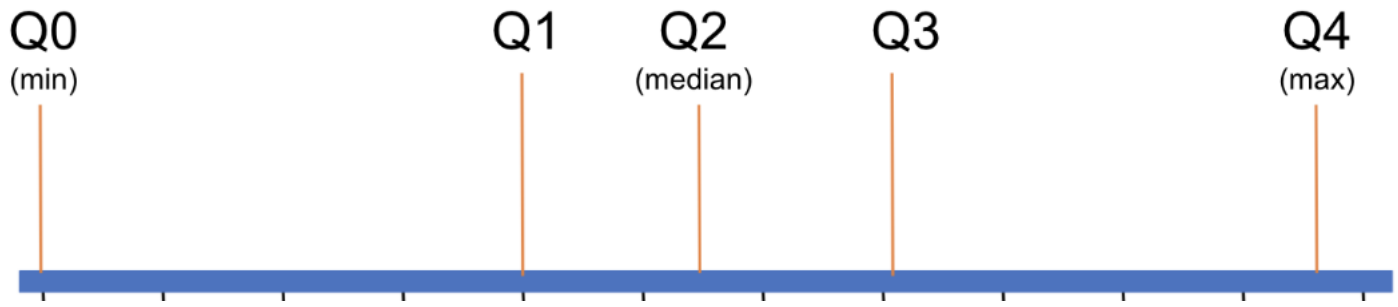
$$\mathrm{CV} = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

- For a sample:

$$\mathrm{CV} = \left(\frac{s}{\bar{x}}\right) \times 100\%$$

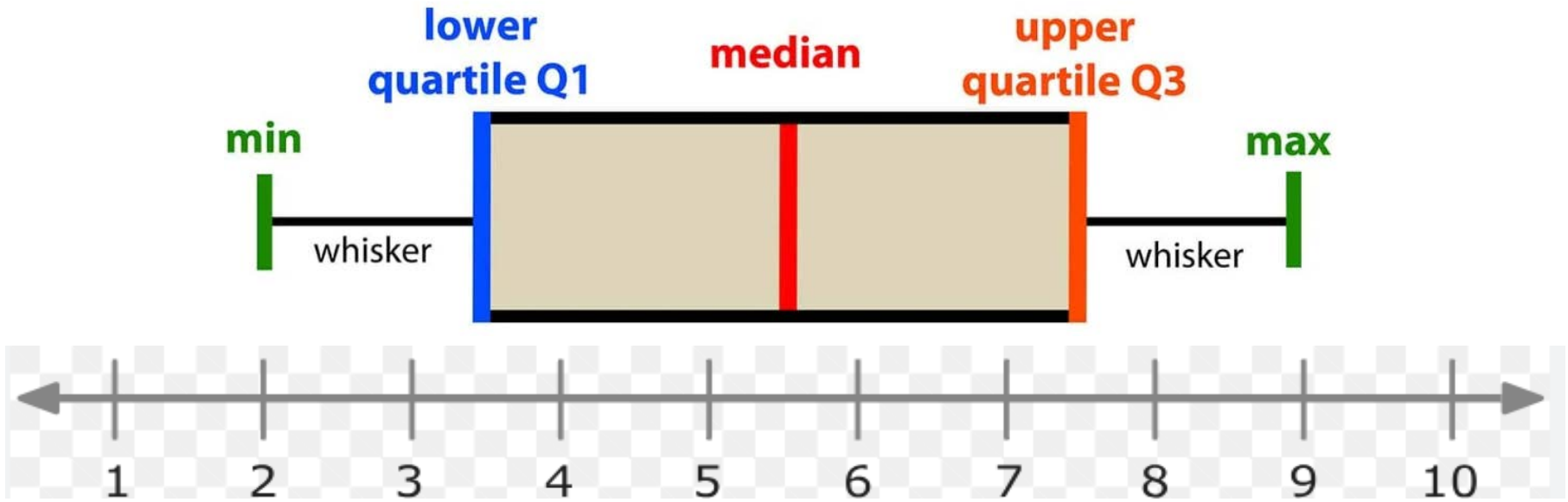
Five-Number Summary in Statistics

The five-number summary is a concise way of summarizing a dataset. It provides a quick overview of the distribution of data points and includes five key descriptive statistics: the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.



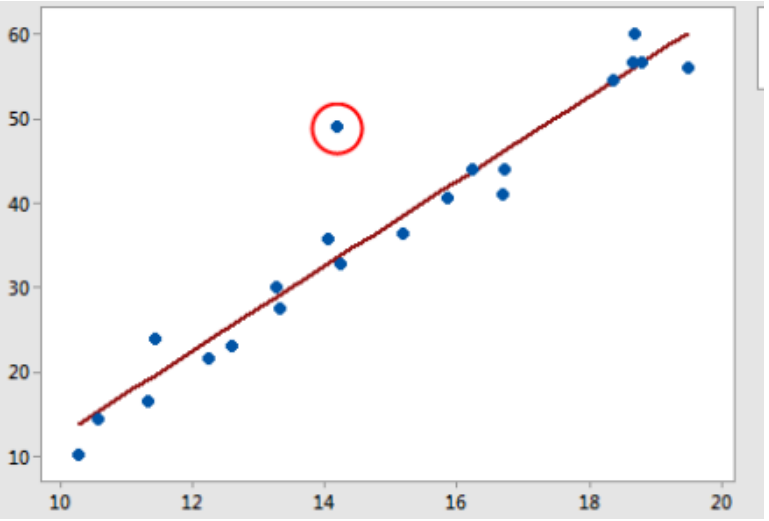
A box and whisker plot

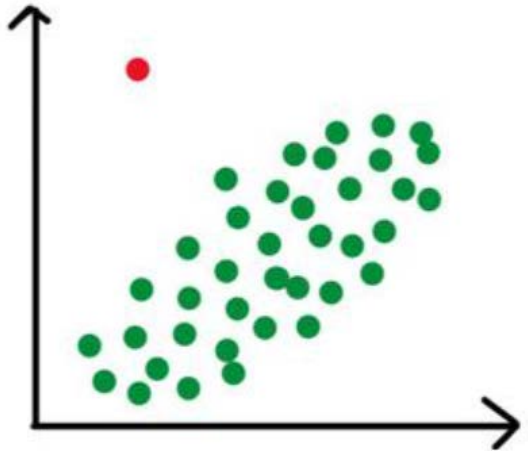
A box and whisker plot—also called a box plot—**displays the five-number summary** of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum. This plot helps in identifying the spread, central tendency, and potential outliers in the data.



Outlier in Statistics

An outlier is a data point that differs significantly from other observations in a dataset. Outliers can occur due to variability in the data, measurement errors, or experimental anomalies. They can affect the overall analysis, including the mean and standard deviation, and can provide valuable insights or indicate issues with the data collection process.





Outliers Calculation