Population & Sample

Population: The entire group that is the subject of the study. It is often impractical or impossible to study every individual in a population, especially when the population is large. For example, if you were interested in studying the average height of all people in a country, the population would be the entire population of that country.

Sample: A subset of the population that is selected for the actual study. Since studying an entire population is often not feasible, researchers select a sample from the population. The goal is for the sample to be representative of the population so that the findings from the sample can be generalized to the entire population.

Population & Sample

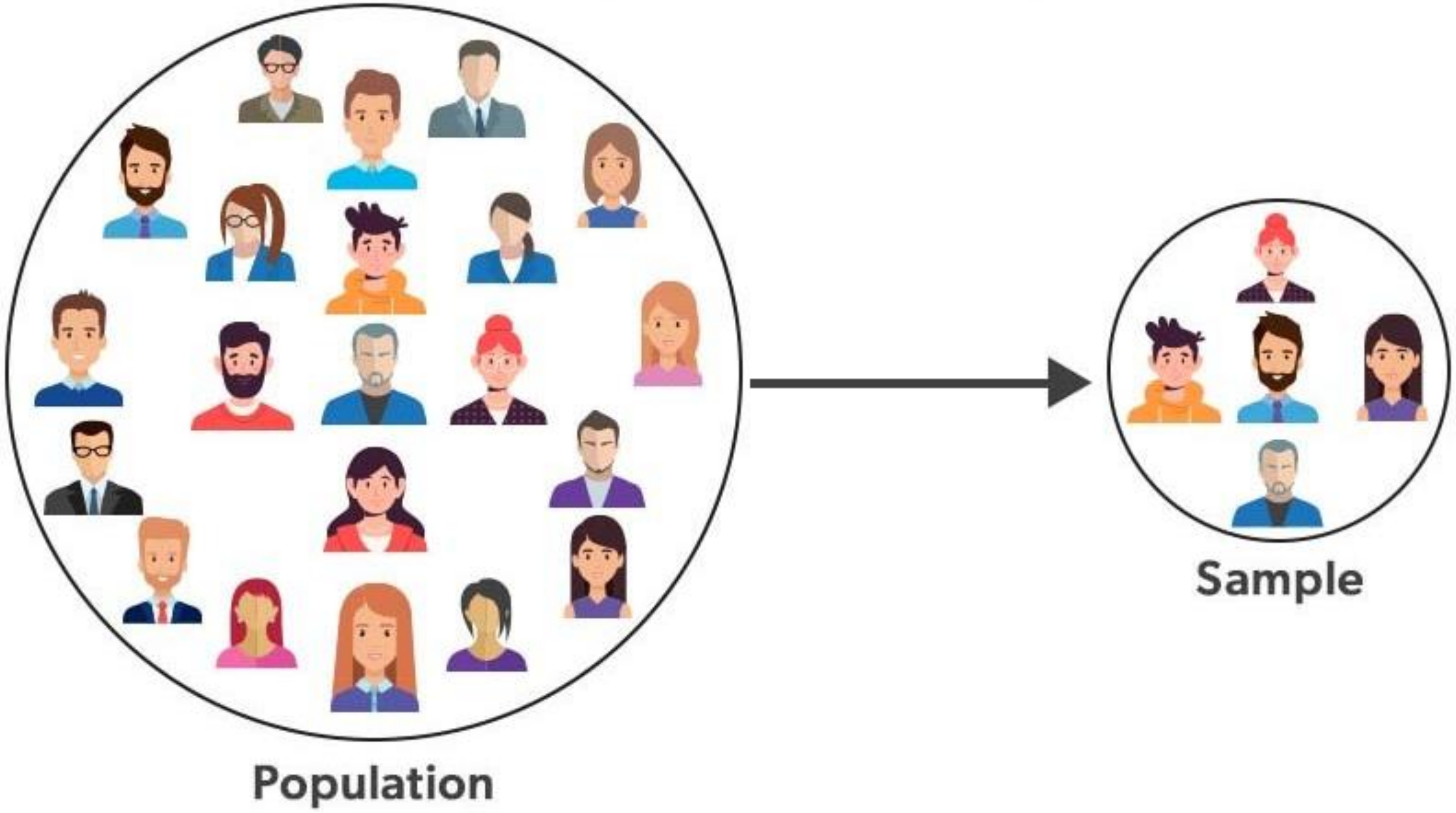
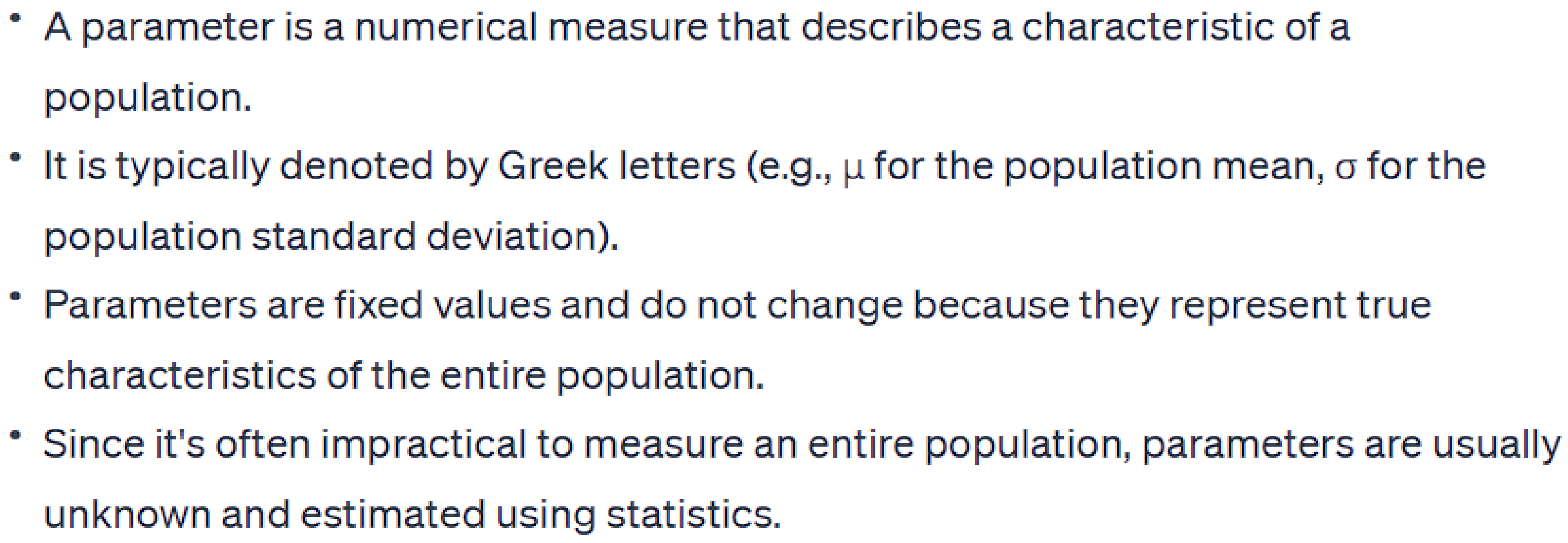


Image source: geeksforgeeks.org

Parameter Vs. Statistic

Parameters and statistics are terms used in statistics to describe numerical measures that summarize or describe the characteristics of a population or a sample.

Parameter:

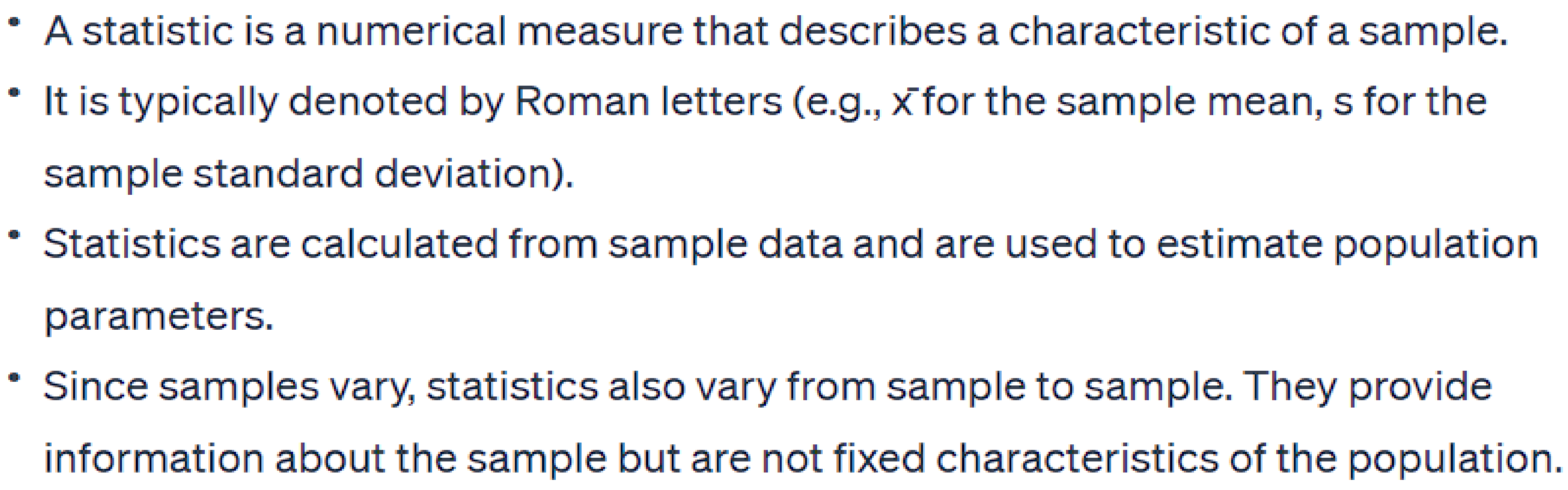


- A parameter is a numerical measure that describes a characteristic of a population.
- It is typically denoted by Greek letters (e.g., μ for the population mean, σ for the population standard deviation).
- Parameters are fixed values and do not change because they represent true characteristics of the entire population.
- Since it's often impractical to measure an entire population, parameters are usually unknown and estimated using statistics.

Parameter Vs. Statistic

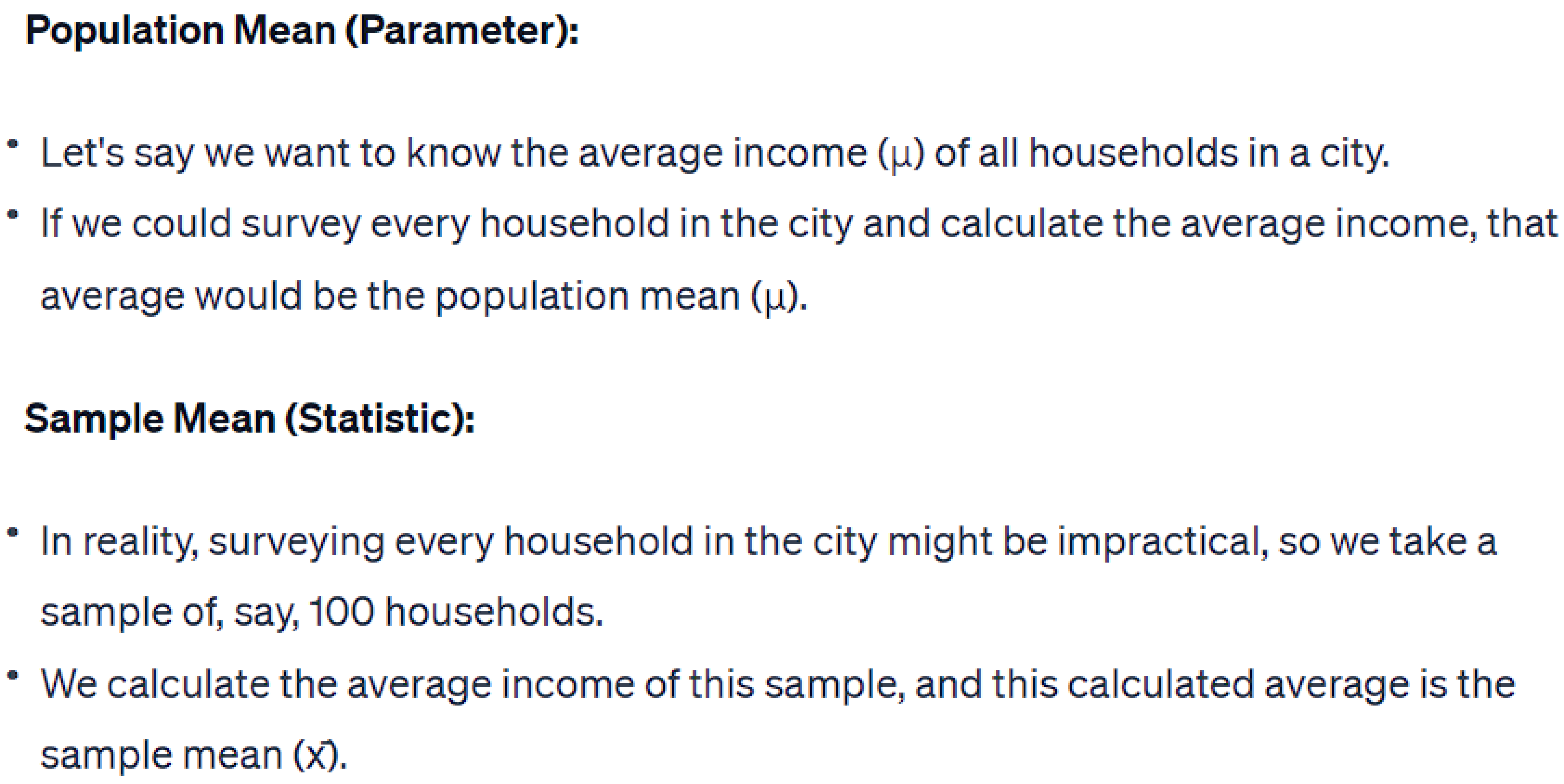
Parameters and statistics are terms used in statistics to describe numerical measures that summarize or describe the characteristics of a population or a sample.

Statistic:



- A statistic is a numerical measure that describes a characteristic of a sample.
- It is typically denoted by Roman letters (e.g., $\bar{x}$ for the sample mean, s for the sample standard deviation).
- Statistics are calculated from sample data and are used to estimate population parameters.
- Since samples vary, statistics also vary from sample to sample. They provide information about the sample but are not fixed characteristics of the population.

Parameter Vs. Statistic



Population Mean (Parameter):

- Let's say we want to know the average income (μ) of all households in a city.
- If we could survey every household in the city and calculate the average income, that average would be the population mean (μ).

Sample Mean (Statistic):

- In reality, surveying every household in the city might be impractical, so we take a sample of, say, 100 households.
- We calculate the average income of this sample, and this calculated average is the sample mean ($\bar{x}$).

Parameter Vs. Statistic

Mathematical Representation:

- Let X be the random variable representing income.
- The population mean (μ) is the average income of the entire population: $\mu = \frac{\sum X}{N}$, where N is the population size.
- The sample mean ($\bar{x}$) is the average income of the sample: $\bar{X} = \frac{\sum x}{n}$, where n is the sample size.

Parameter	**Statistic**
μ : population mean	$\bar{x}$: sample mean
σ : population standard deviation	s : sample standard deviation
p : population proportion	$\hat{p}$: sample proportion

Measures of Central Tendency

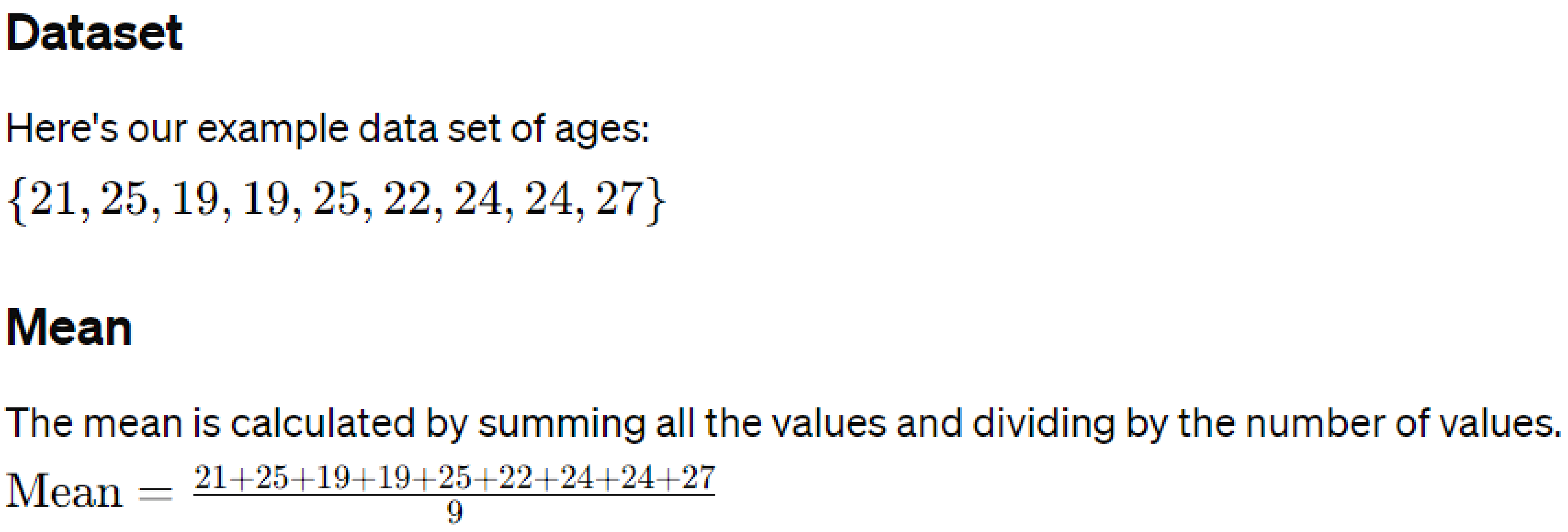
Measures of central tendency are statistical indicators that describe the central position of a frequency distribution for a data set.

- **Mean**: The average of all data points. It is calculated by summing all the values in the data set and then dividing by the number of values.

- **Median**: The middle value in the data set when it is arranged in ascending or descending order. If the number of observations is even, the median is the average of the two middle numbers.

- **Mode**: The value that appears most frequently in the data set. A data set can have one mode (unimodal), two modes (bimodal), or more modes (multimodal).

Measures of Central Tendency



Dataset

Here's our example data set of ages:

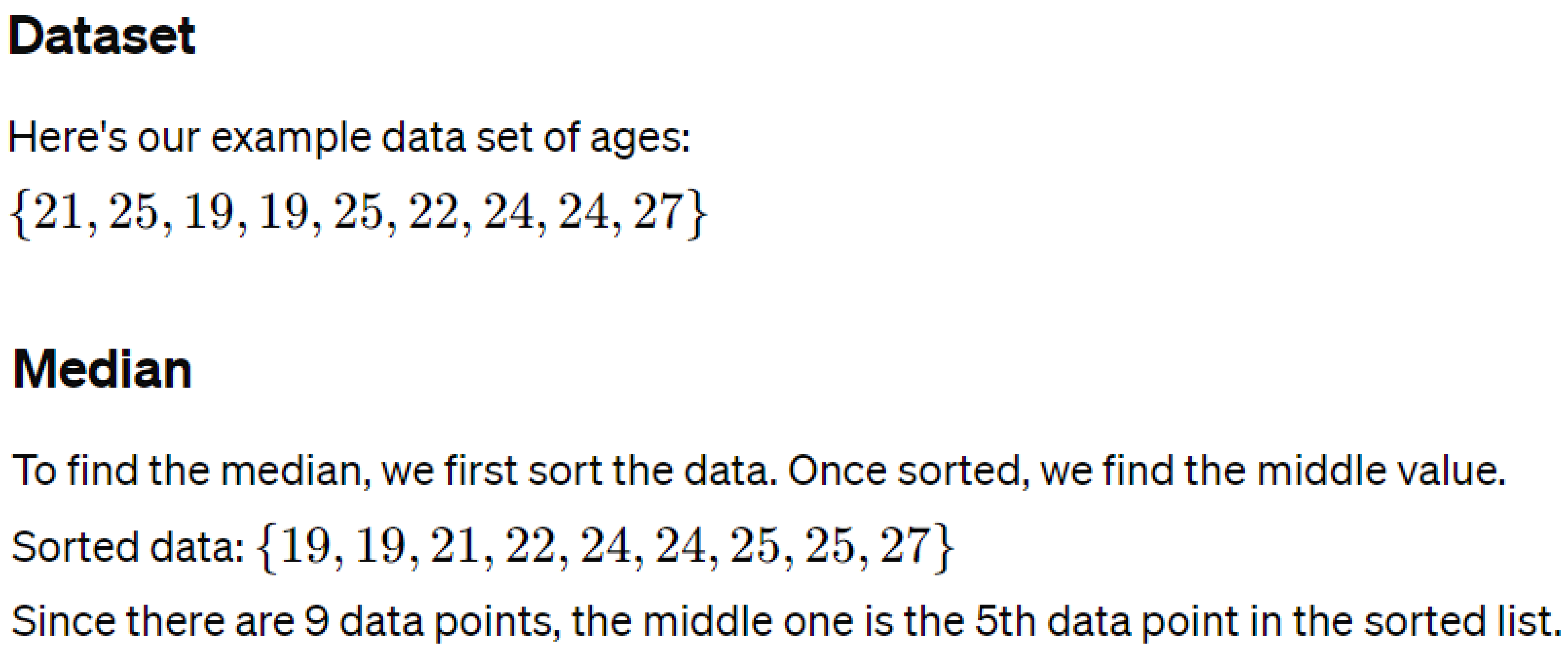
$$\{21, 25, 19, 19, 25, 22, 24, 24, 27\}$$

Mean

The mean is calculated by summing all the values and dividing by the number of values.

$$\text{Mean} = \frac{21+25+19+19+25+22+24+24+27}{9}$$

Measures of Central Tendency



Dataset

Here's our example data set of ages:

$\{21, 25, 19, 19, 25, 22, 24, 24, 27\}$

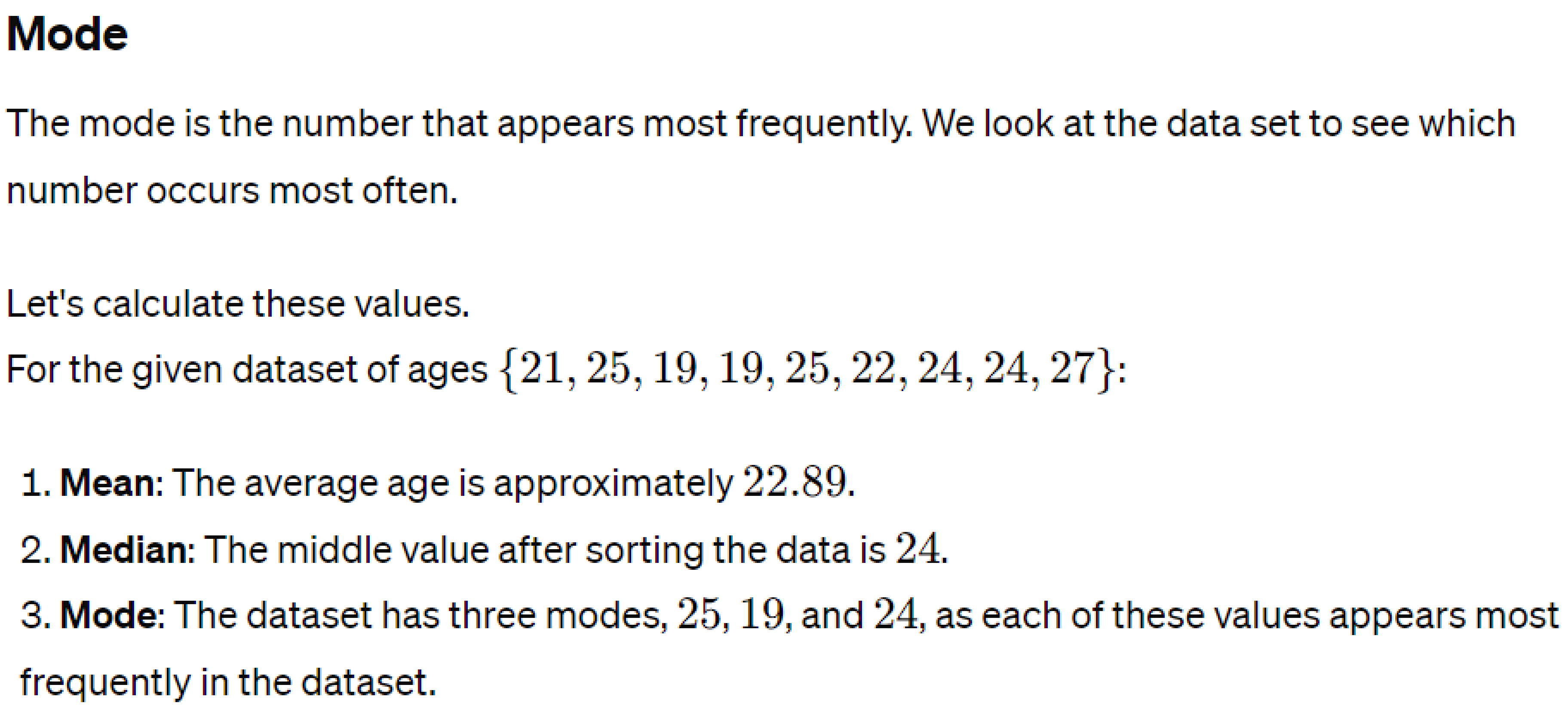
Median

To find the median, we first sort the data. Once sorted, we find the middle value.

Sorted data: $\{19, 19, 21, 22, 24, 24, 25, 25, 27\}$

Since there are 9 data points, the middle one is the 5th data point in the sorted list.

Measures of Central Tendency



Mode

The mode is the number that appears most frequently. We look at the data set to see which number occurs most often.

Let's calculate these values.
For the given dataset of ages $\{21, 25, 19, 19, 25, 22, 24, 24, 27\}$:

1. **Mean**: The average age is approximately 22.89.
2. **Median**: The middle value after sorting the data is 24.
3. **Mode**: The dataset has three modes, 25, 19, and 24, as each of these values appears most frequently in the dataset.

Standard Deviation

Standard deviation is a statistical measure that quantifies the amount of variation or dispersion of a set of values.

- If the data points are close to the mean, then the standard deviation is small; this indicates that the data points tend to be close to the mean of the data set.

- Conversely, if the data points are widely spread out from the mean, the standard deviation is large; this shows that there is a greater level of variance or uncertainty in the data set.

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

Fig: STD for Population

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

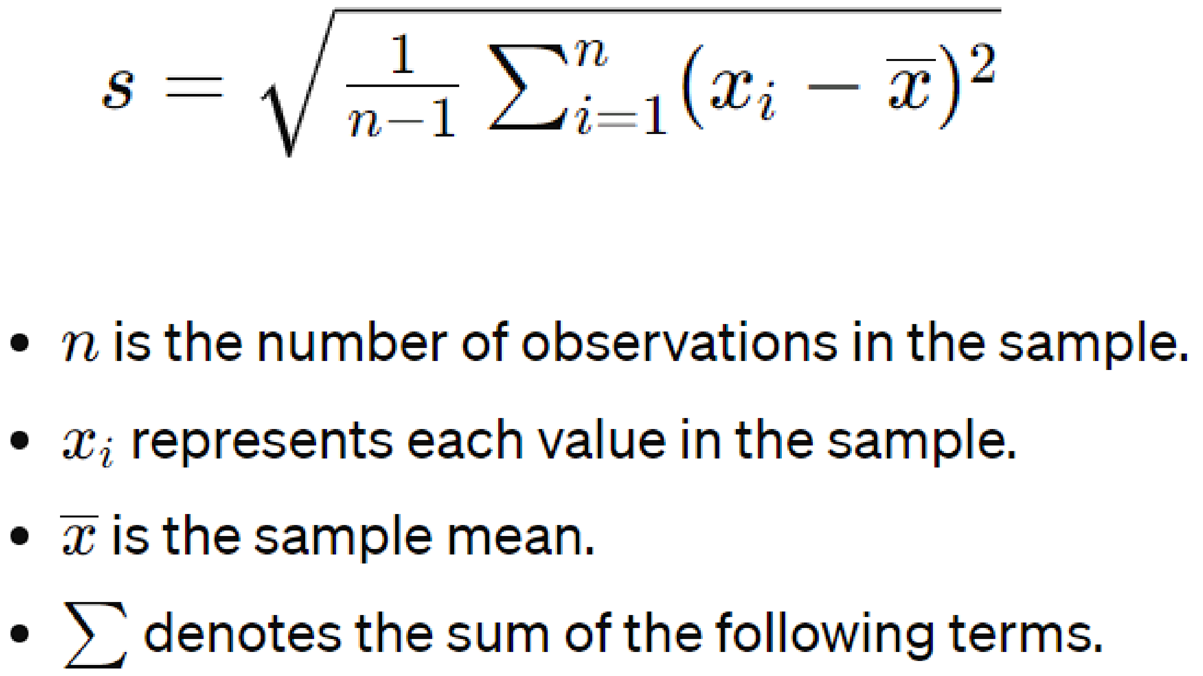
Fig: STD for Sample

Standard Deviation

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

- σ (sigma) is the symbol used for the population standard deviation.
- N is the total number of observations in the population.
- x_i represents each individual value in the population.
- μ (mu) is the population mean (average).

Standard Deviation

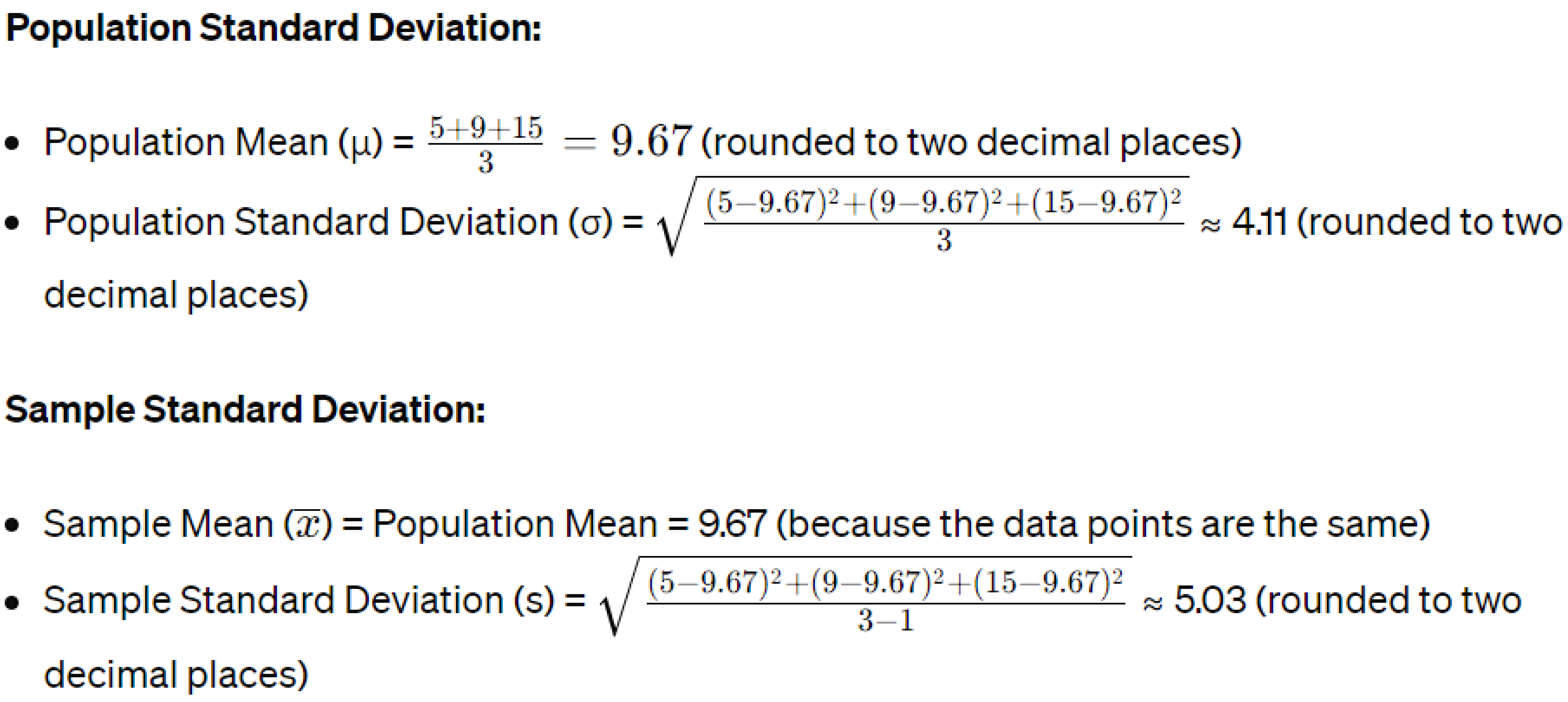


$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

- n is the number of observations in the sample.
- x_i represents each value in the sample.
- $\overline{x}$ is the sample mean.
- $\sum$ denotes the sum of the following terms.

Standard Deviation

Let’s calculate for dataset = [5, 9, and 15]



Population Standard Deviation:

- Population Mean (μ) = $\frac{5+9+15}{3} = 9.67$ (rounded to two decimal places)
- Population Standard Deviation (σ) = $\sqrt{\frac{(5-9.67)^2+(9-9.67)^2+(15-9.67)^2}{3}} \approx$ 4.11 (rounded to two decimal places)

Sample Standard Deviation:

- Sample Mean ($\overline{x}$) = Population Mean = 9.67 (because the data points are the same)
- Sample Standard Deviation (s) = $\sqrt{\frac{(5-9.67)^2+(9-9.67)^2+(15-9.67)^2}{3-1}} \approx$ 5.03 (rounded to two decimal places)

Standard Deviation

Importance of Standard Deviation:

- Quantifying Variability: Standard deviation measures how much data points differ from the average, offering a clear picture of the spread.
- Risk Assessment: In finance and investments, it gauges risk by indicating how much returns can deviate from the expected average.
- Quality Control: It's used in manufacturing to determine the consistency of product quality and identify defects.
- Comparing Datasets: Standard deviation enables the comparison of variability between different sets of data, even with different means.
- Statistical Inference: It's essential for hypothesis testing and determining the significance of results in scientific research.

Variance

Variance measures the average degree to which each point differs from the mean. If all the numbers are the same, the variance is 0, because every number is equal to the mean.

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

Fig: Var for Population

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

Fig: Var for Sample

Variance

Standard deviation and variance are closely related statistical measures used to quantify the spread of data points in a dataset around the mean. Here are the main differences between them:

- **Units:**
  - **Variance:** The units of variance are the squares of the units of the data points. For example, if the data is measured in meters, the variance will be in square meters (m^2).
  - **Standard Deviation:** The units of standard deviation are the same as the units of the data points themselves. If the data is in meters, then the standard deviation will also be in meters. This makes the standard deviation more interpretable in the context of the original data.
- **Interpretation:**
  - **Variance:** Provides a measure of the data's spread based on the squared deviations, which emphasizes larger deviations due to the squaring process. It's less intuitive because it does not express variability in the same units as the data.
  - **Standard Deviation:** By taking the square root of the variance, the standard deviation provides a more interpretable measure of spread that is in the same units as the data. This helps in understanding how much variation exists from the average.

Variance

❑ **Mathematical Properties:**

- ❖ **Variance:** Easier to handle in mathematical formulas, especially in statistical inference and other calculations, because it simplifies the math when dealing with squared terms.

- ❖ **Standard Deviation:** Although more intuitive, it can be more complex to use directly in some types of statistical calculations due to the square root.

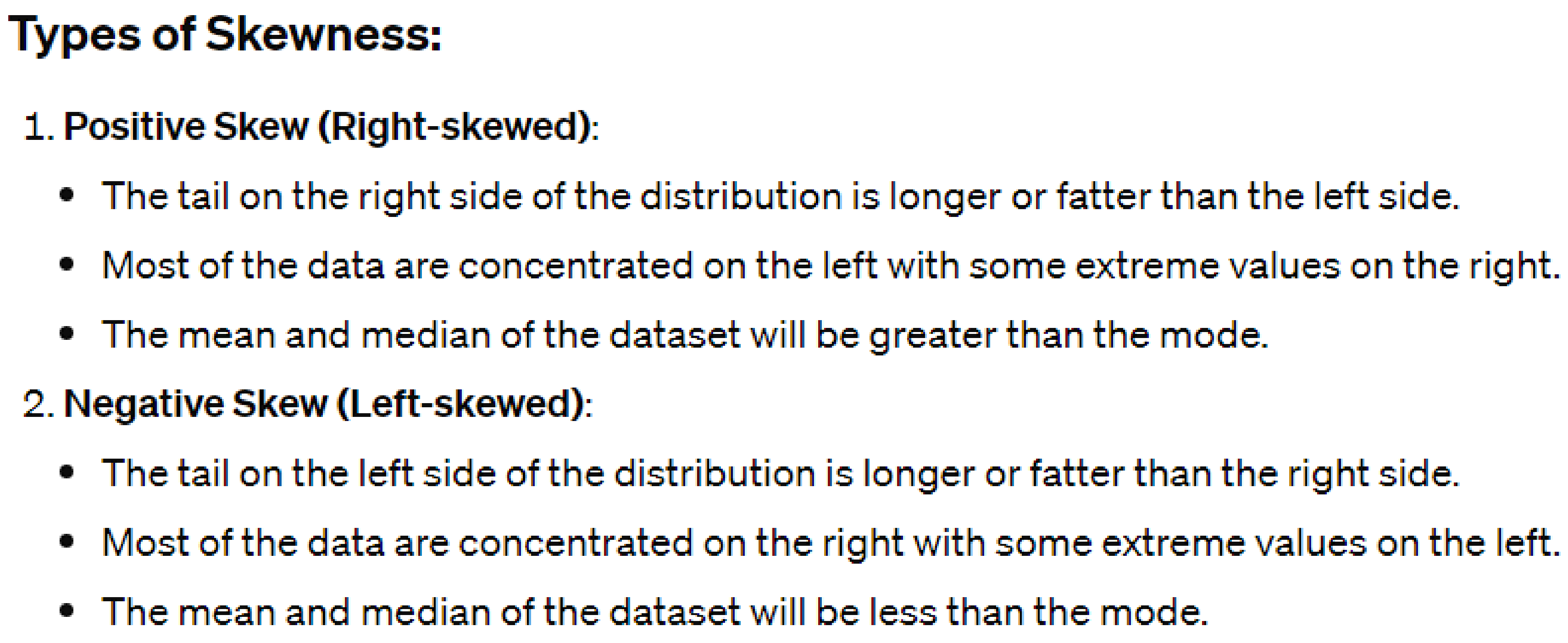
❑ **Sensitivity:**

- ❖ **Variance:** Since it squares the deviations from the mean, variance is more sensitive to outliers and large deviations. Outliers can have a disproportionately large effect on the variance.

- ❖ **Standard Deviation:** While also sensitive to outliers, its real-world interpretation as the average distance from the mean can be easier for typical data analysis purposes.

Skewness in Statistics

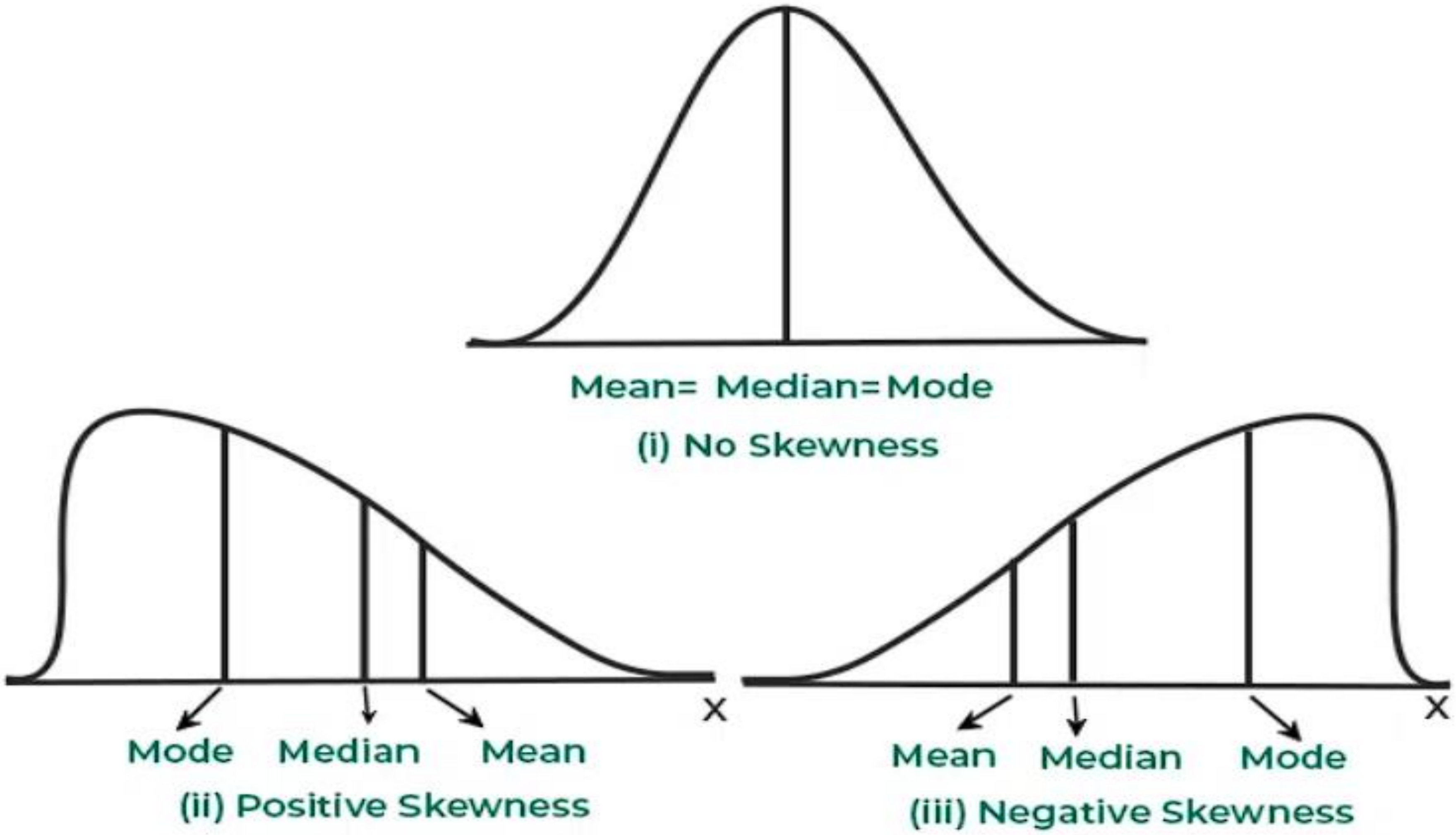
Skewness is a statistical measure that describes the asymmetry of a data distribution around its mean. It indicates whether the data are spread out more to one side of the mean or the other, essentially measuring the lack of symmetry in the data.



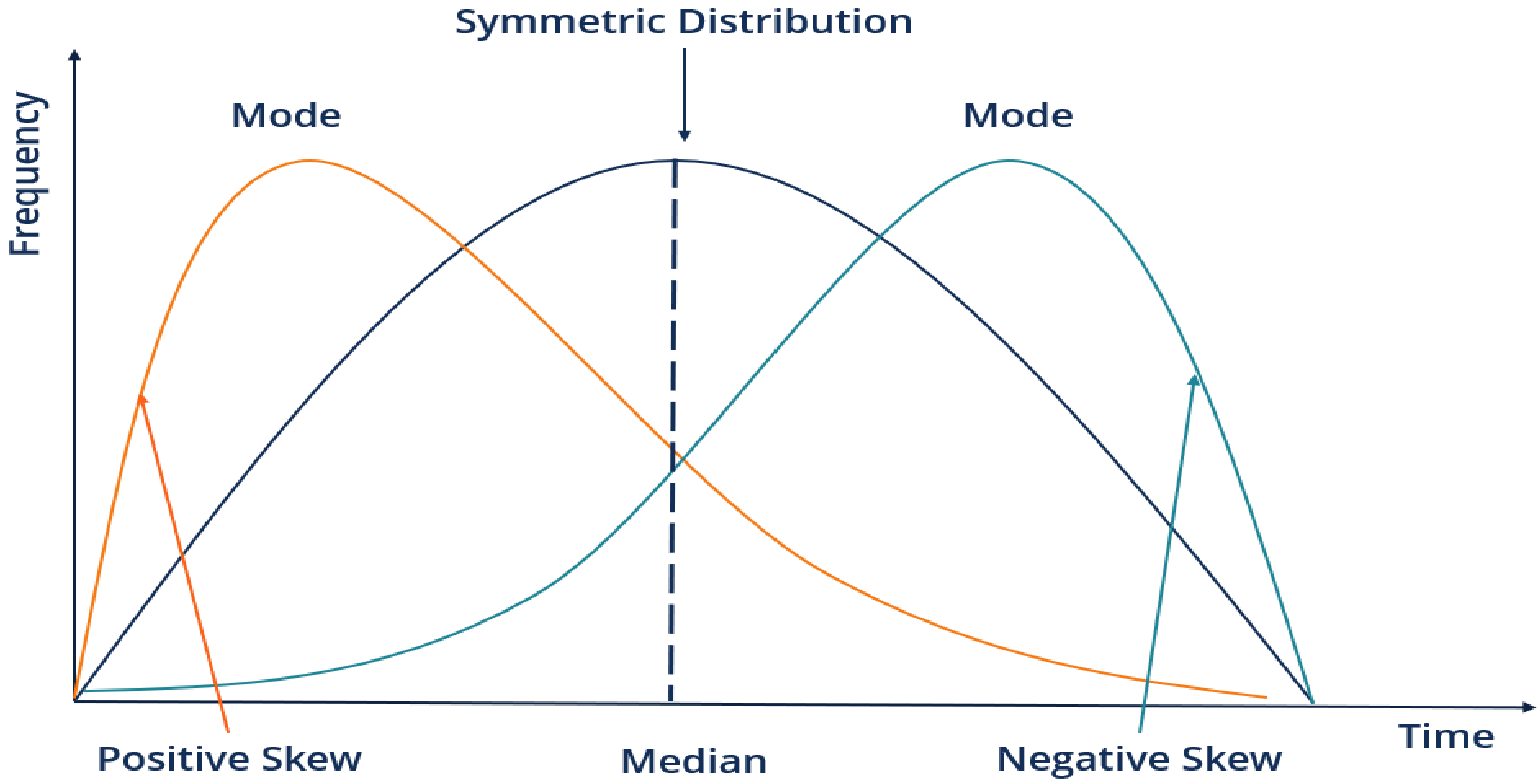
Types of Skewness:

1. **Positive Skew (Right-skewed)**:
   - The tail on the right side of the distribution is longer or fatter than the left side.
   - Most of the data are concentrated on the left with some extreme values on the right.
   - The mean and median of the dataset will be greater than the mode.
2. **Negative Skew (Left-skewed)**:
   - The tail on the left side of the distribution is longer or fatter than the right side.
   - Most of the data are concentrated on the right with some extreme values on the left.
   - The mean and median of the dataset will be less than the mode.

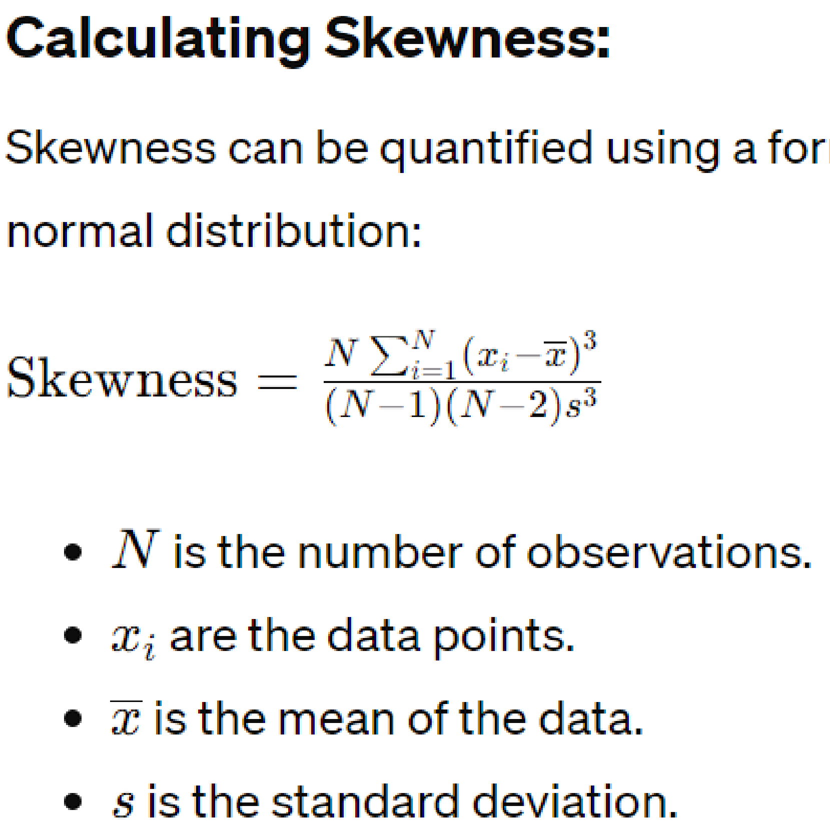
Skewness in Statistics



Skewness in Statistics



Skewness in Statistics



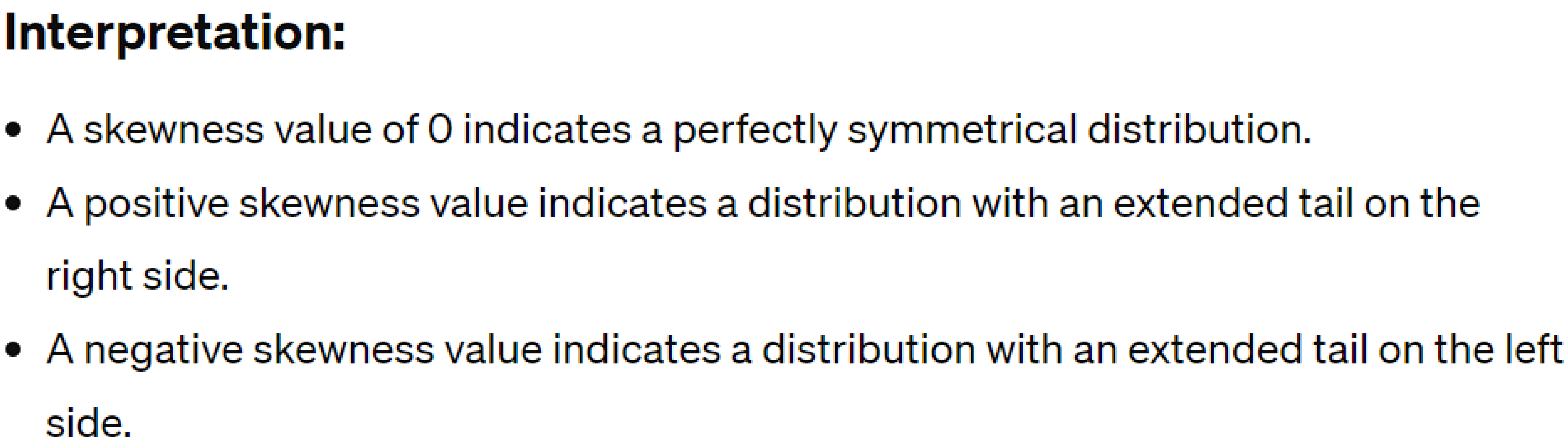
Calculating Skewness:

Skewness can be quantified using a formula that compares the spread of data to a normal distribution:

$$\text{Skewness} = \frac{N\sum_{i=1}^{N}(x_i-\overline{x})^3}{(N-1)(N-2)s^3}$$

- N is the number of observations.
- x_i are the data points.
- $\overline{x}$ is the mean of the data.
- s is the standard deviation.

Skewness in Statistics



Interpretation:

- A skewness value of 0 indicates a perfectly symmetrical distribution.
- A positive skewness value indicates a distribution with an extended tail on the right side.
- A negative skewness value indicates a distribution with an extended tail on the left side.

Skewness in Statistics

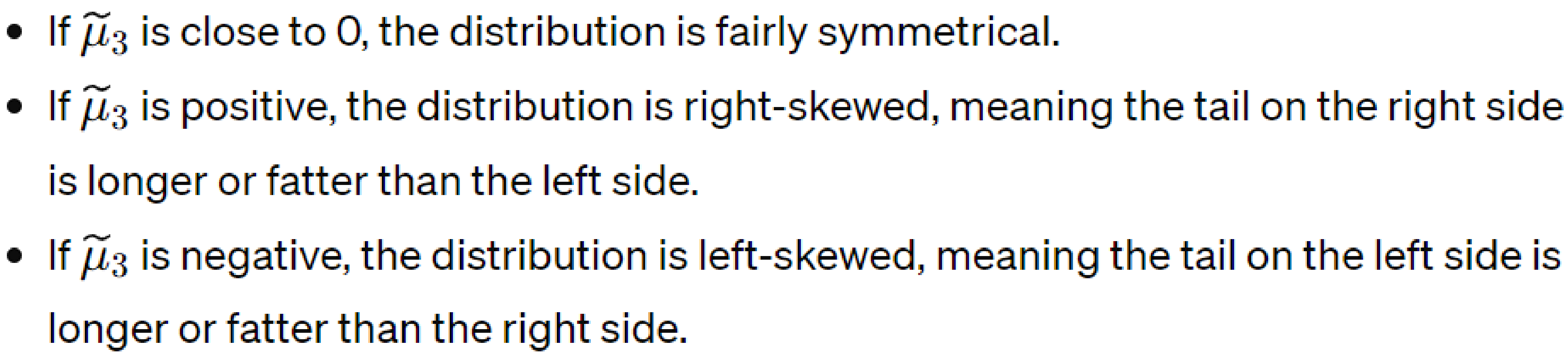
Another Approach:

$$\tilde{\mu}_3 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^3}{(N-1)\cdot\sigma^3}$$

- $\tilde{\mu}_3$ is the skewness of the distribution.
- N is the number of variables (or observations) in the distribution.
- X_i is each individual random variable (data point).
- $\overline{X}$ is the mean (average) of the distribution.
- σ is the standard deviation of the distribution.

Skewness in Statistics

Interpretation:

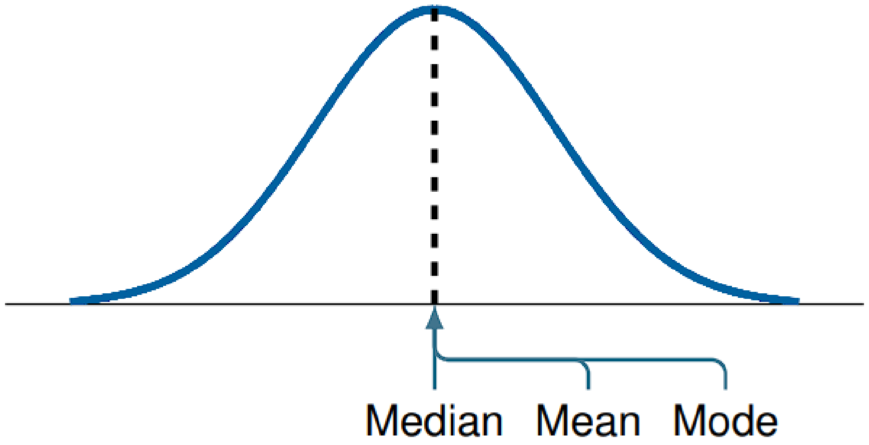


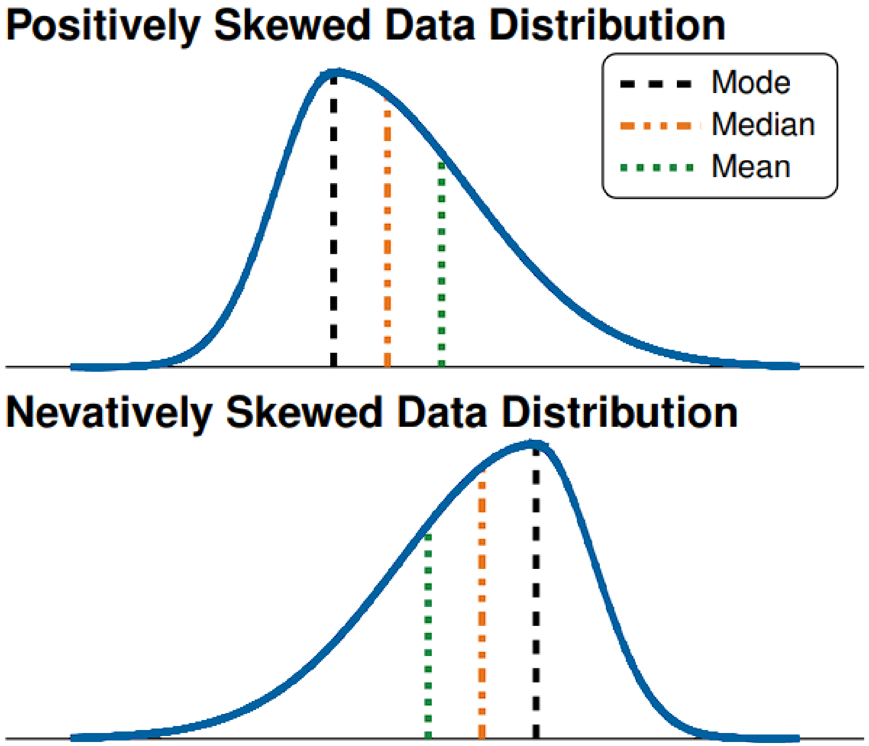
- If $\widetilde{\mu}_3$ is close to 0, the distribution is fairly symmetrical.
- If $\widetilde{\mu}_3$ is positive, the distribution is right-skewed, meaning the tail on the right side is longer or fatter than the left side.
- If $\widetilde{\mu}_3$ is negative, the distribution is left-skewed, meaning the tail on the left side is longer or fatter than the right side.

Normal Distribution

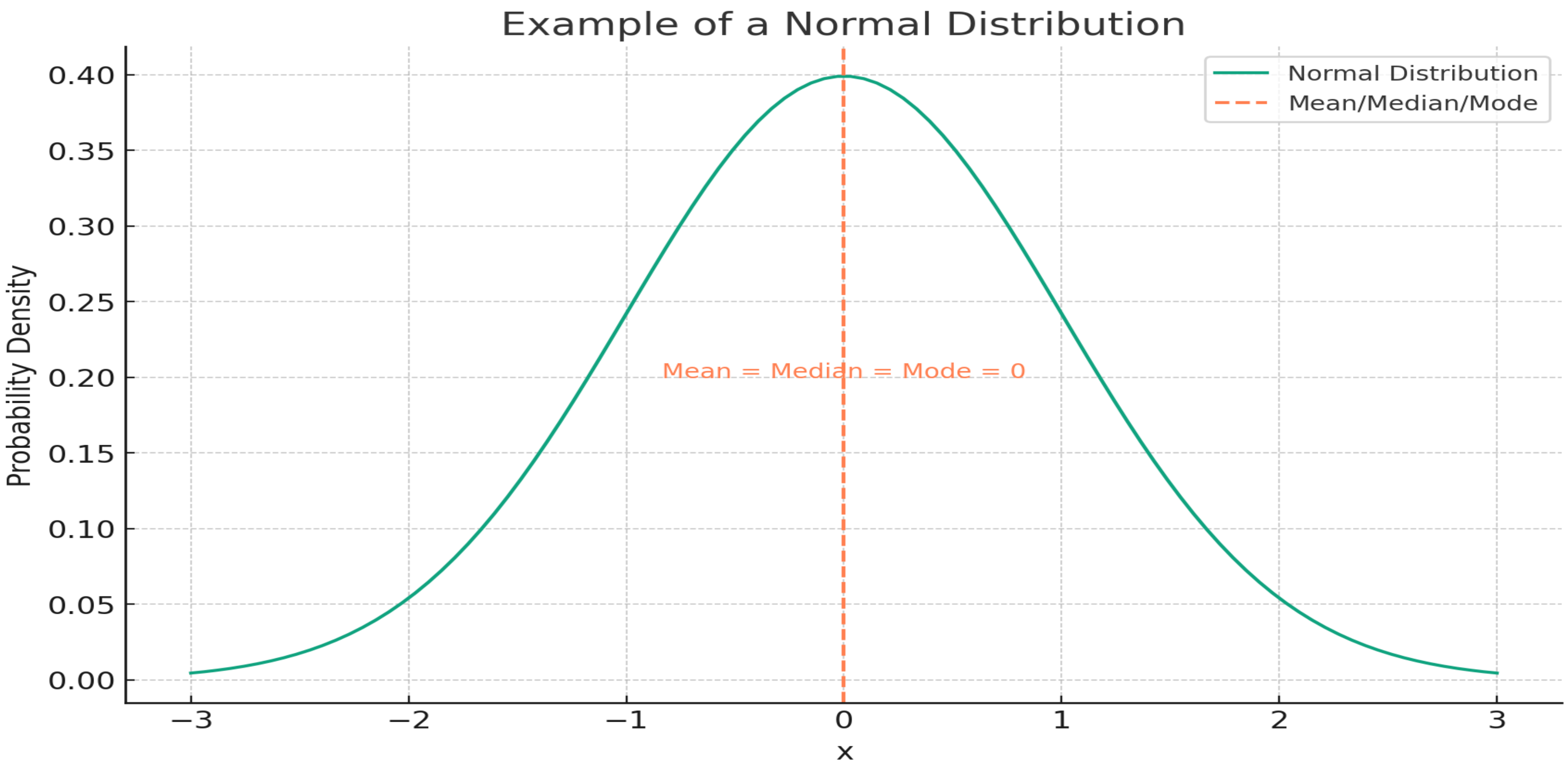
Normal distribution

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



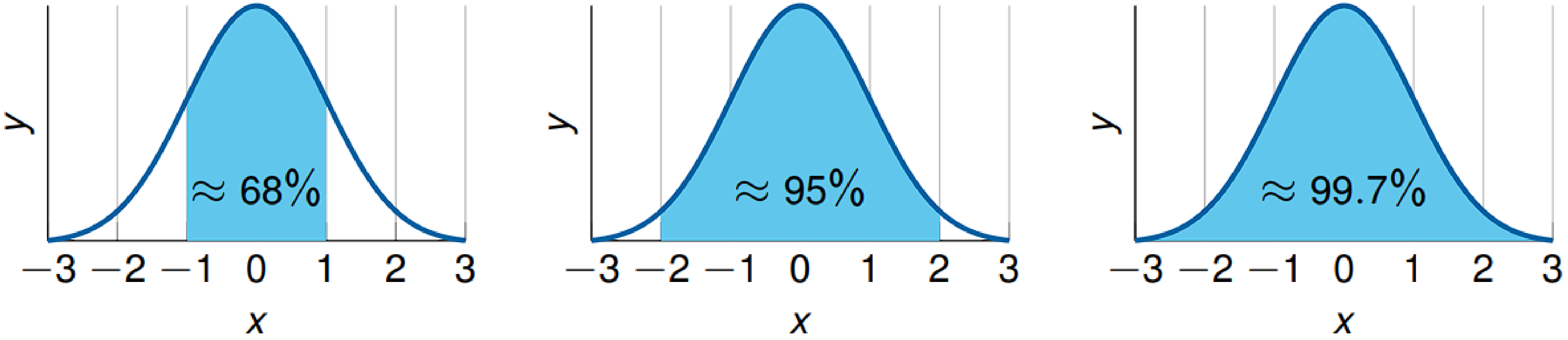


Normal Distribution



Normal Distribution

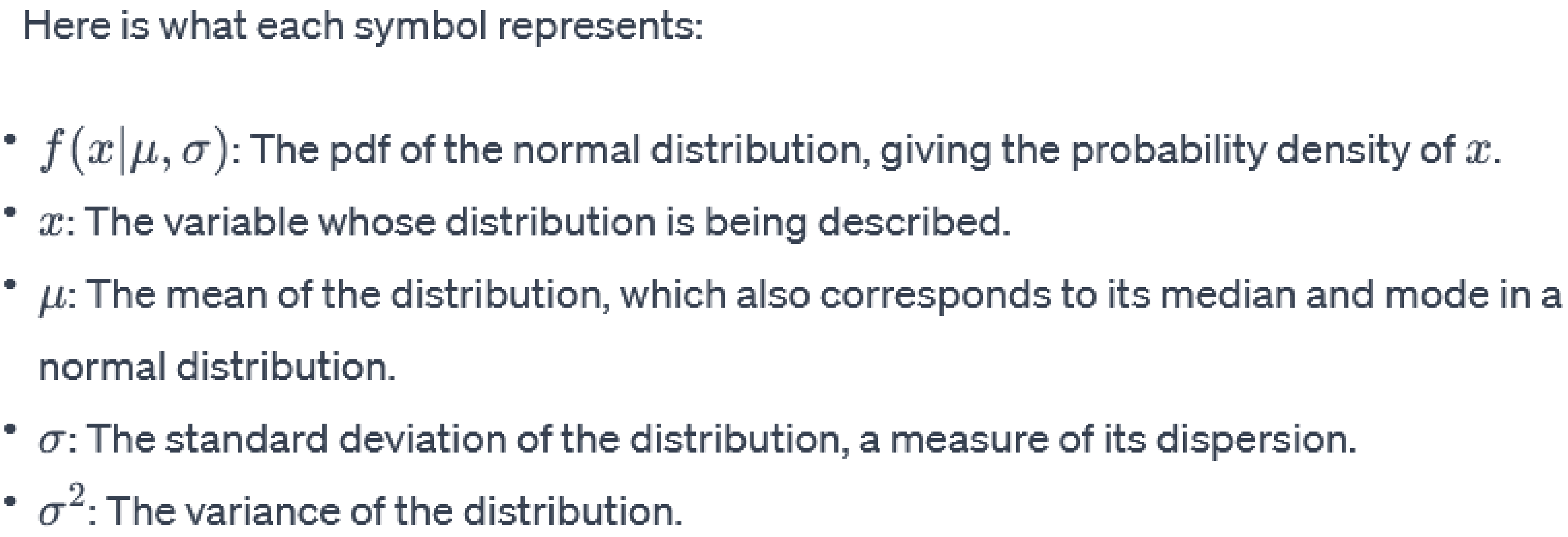
- **The normal distribution:**
  - From $\mu - \sigma$ to $\mu + \sigma$: contains about 68% of the measurements.
    - μ: mean,
    - σ: standard deviation.
  - From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95% of the surface under the curve.
  - $\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7% of the surface under the curve.



Normal Distribution

Probability density function (pdf) for the normal distribution:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



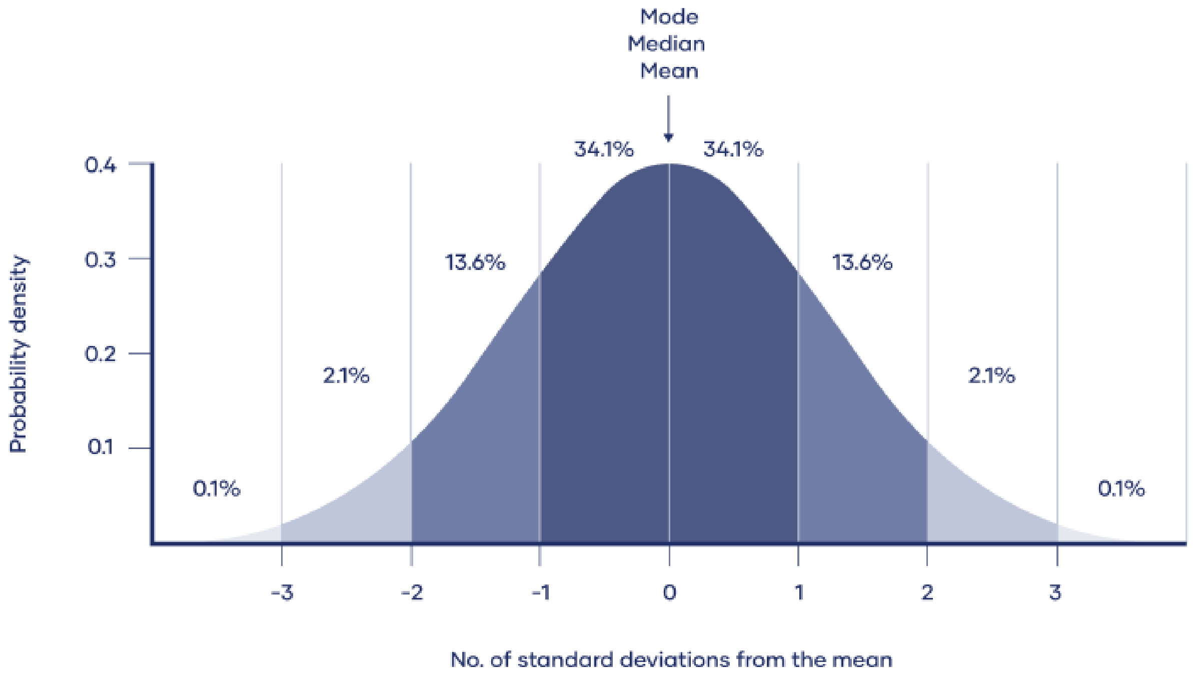
Here is what each symbol represents:

- $f(x|\mu,\sigma)$: The pdf of the normal distribution, giving the probability density of x.
- x: The variable whose distribution is being described.
- μ: The mean of the distribution, which also corresponds to its median and mode in a normal distribution.
- σ: The standard deviation of the distribution, a measure of its dispersion.
- σ^2: The variance of the distribution.

Normal Distribution

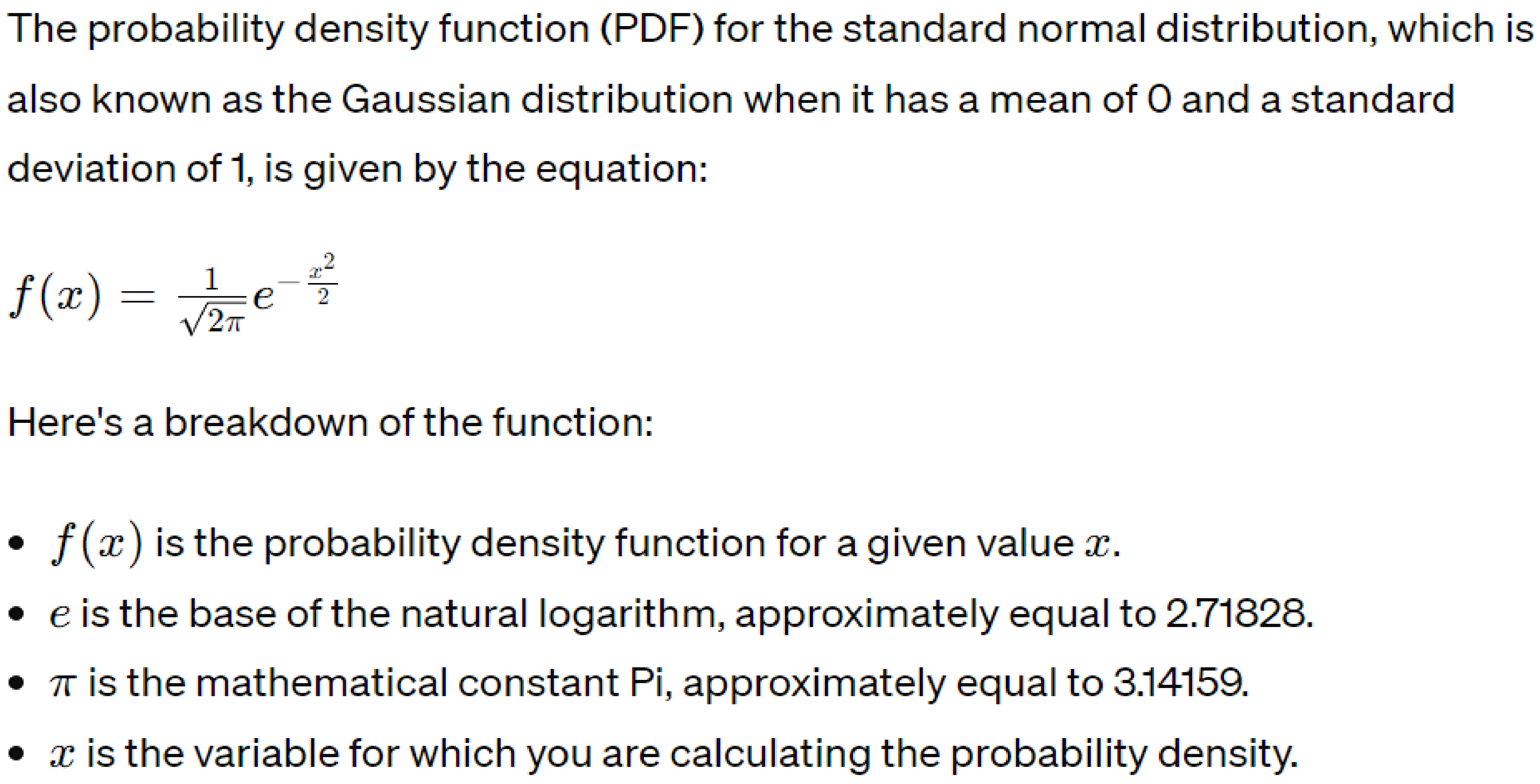
Properties of Normal Distribution:

- The mean, median, and mode are the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.



- Income Distribution In Economy
- Shoe Size
- Birth Weight
- Spending Days in Hospital

Standard Normal Distribution



The probability density function (PDF) for the standard normal distribution, which is also known as the Gaussian distribution when it has a mean of 0 and a standard deviation of 1, is given by the equation:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Here's a breakdown of the function:

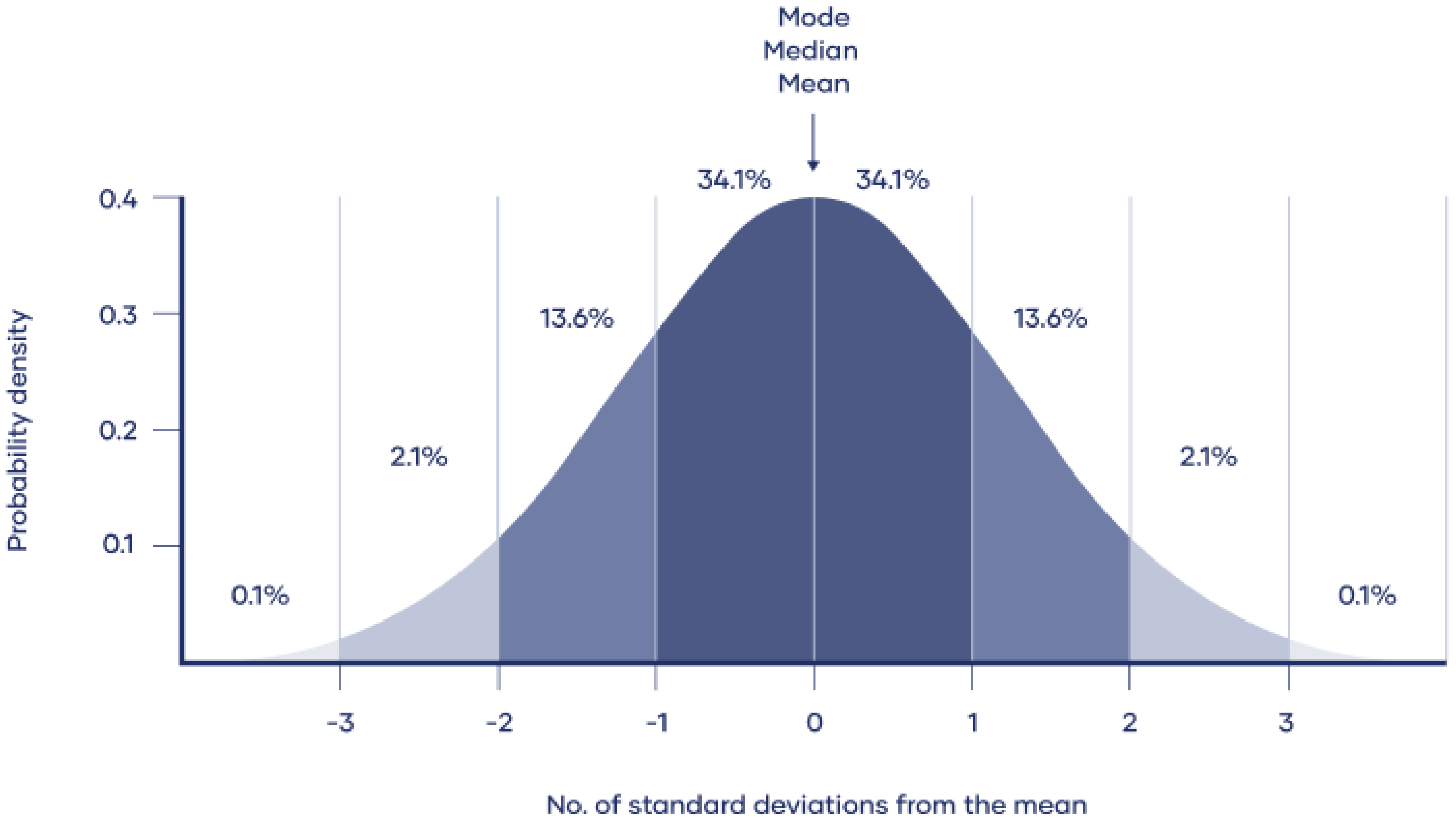
- $f(x)$ is the probability density function for a given value x.
- e is the base of the natural logarithm, approximately equal to 2.71828.
- π is the mathematical constant Pi, approximately equal to 3.14159.
- x is the variable for which you are calculating the probability density.

Z-Score

How to Calculate a Z-Score,

$$Z = \frac{\overset{\text{Score}}{x} - \overset{\text{Mean}}{\mu}}{\underset{\text{SD}}{\sigma}}$$

Z-Score



Mode
Median
Mean
34.1%
34.1%
0.4
0.3
0.2
0.1
Probability density
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
-3
-2
-1
0
1
2
3
No. of standard deviations from the mean

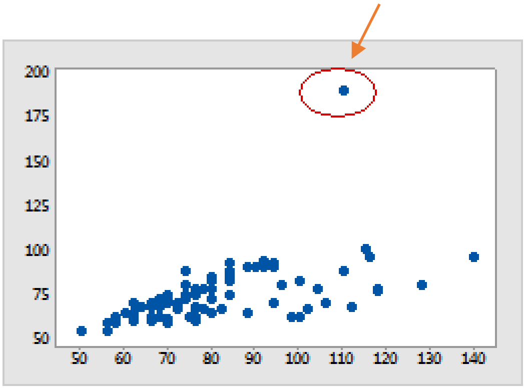
Outliers/ Anomalies Concepts

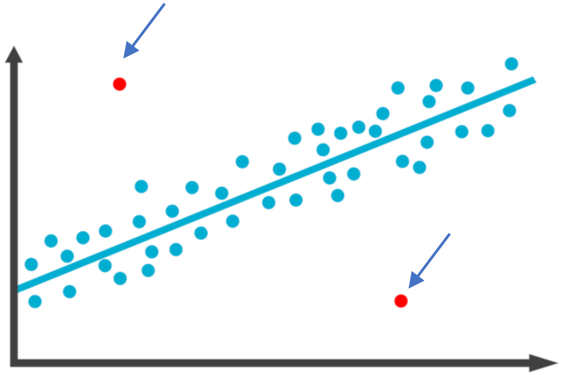
Outliers are data points that are **unusually far** from the central tendency of a distribution.

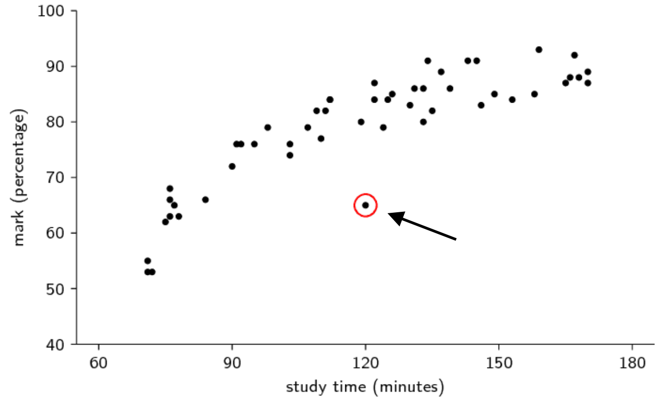
Outliers can occur for various reasons, including errors in **data collection**, **measurement variability**, or **genuine anomalies** in the data. Identifying and handling outliers is important in statistical analysis because they can have a significant impact on the results and interpretation of statistical measures.

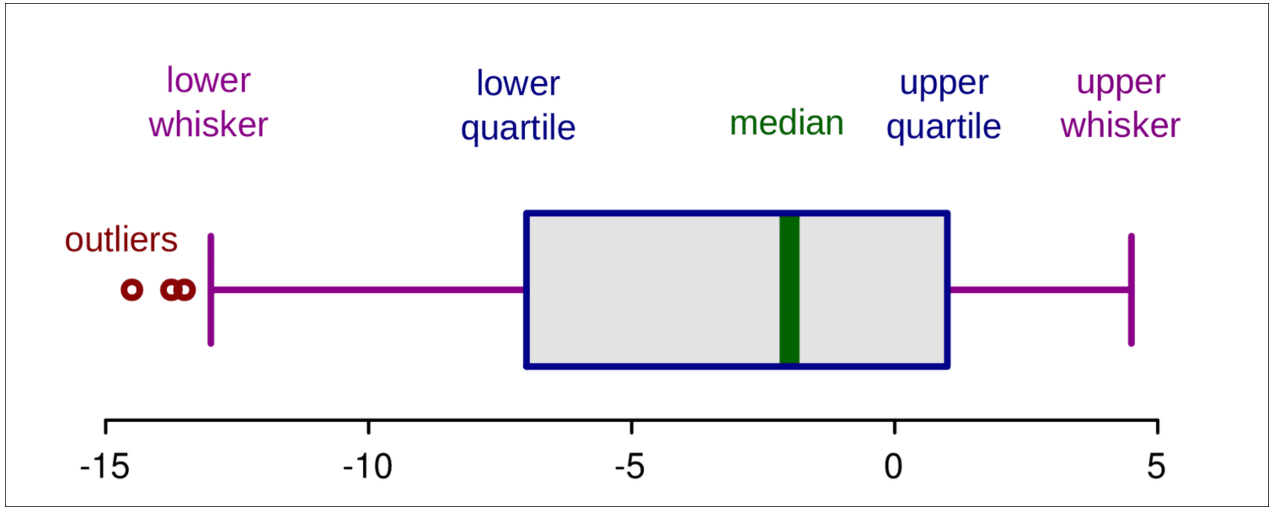
Outliers are a specific subset of anomalies. All outliers are anomalies, but not all anomalies are necessarily outliers.

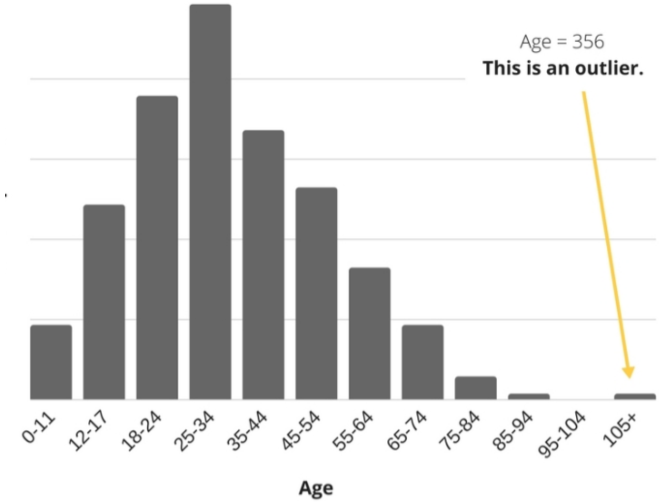
Outliers/ Anomalies Concepts











Outliers/ Anomalies Concepts

1.Causes of Anomalies:

Anomalies can be caused by errors in data collection, measurement inaccuracies, genuine rare events, or changes in the underlying processes generating the data.

2.Types of Anomalies:

Anomalies can be broadly categorized into three types:

1. **Point Anomalies:** Individual data points that are significantly different from the rest.
2. **Contextual Anomalies:** Anomalous Data points within a specific context or subpopulation.
3. **Collective Anomalies:** Groups of data points that exhibit anomalous behaviour when considered together.

Outliers/ Anomalies Concepts

Here are key points for detecting anomalies in machine learning:

1. Statistical Methods:
- Use z-score, modified z-score, or percentile scores to identify data points with extreme values.

2. Distance-Based Methods:
- Calculate Euclidean or Mahalanobis distances to measure deviations from the mean or centroid.

3. Density-Based Methods:
- Apply DBSCAN, Isolation Forest, or One-Class SVM to identify anomalies based on data density.

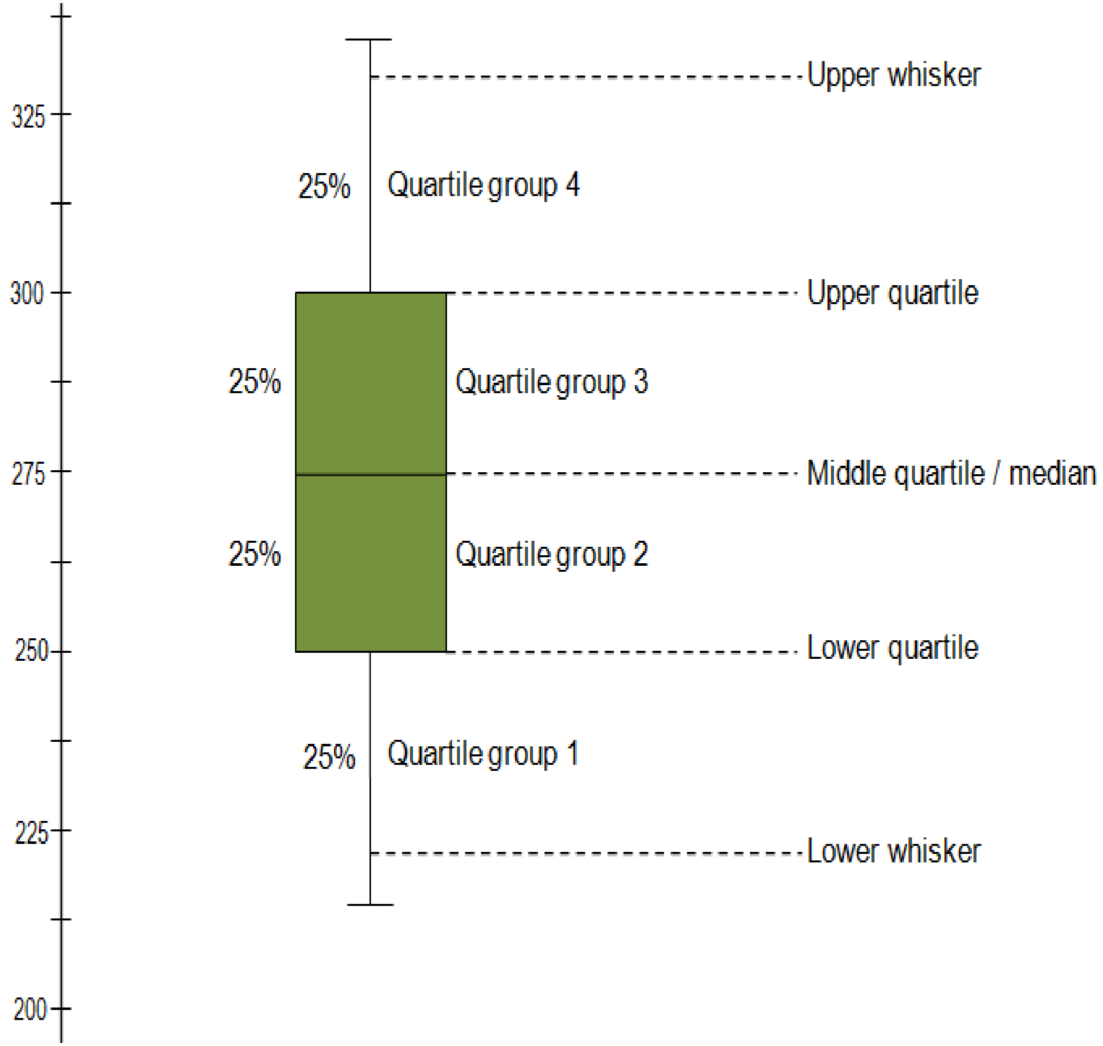
4. Clustering Methods:
- Employ K-Means or hierarchical clustering to identify sparse clusters as anomalies.

5. Machine Learning Models:
- Use Autoencoders or One-Class SVM for anomaly detection based on learning normal patterns.

6. Time Series Methods:
- Apply ARIMA, Prophet, or other time series models to detect anomalies in time-dependent data.

Boxplot for Outliers Detection

Typically, any data point that falls below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR is considered an outlier. This is a general guideline, and the threshold of 1.5 can be adjusted based on the specific requirements of your analysis or the characteristics of your dataset.



Boxplot for Outliers Detection

The first quartile (Q1) corresponds to the 25th percentile, which means 25% of the data falls below Q1. The second quartile (Q2) is the median and corresponds to the 50th percentile, indicating that 50% of the data falls below Q2. The third quartile (Q3) corresponds to the 75th percentile, where 75% of the data falls below Q3.

- 1st Quartile (Q1) or 25 percentile = ((25/100) * (N+1)) = P_th index
- Q2 = Median
- 3rd Quartile (Q3) or 75 percentile = ((75/100) * (N+1))
- IQR = Q3 – Q1

- Lower Whisker = Q1 – (1.5*IQR)
- Upper Whisker = Q3 + (1.5*IQR)

Here,
N = Total Data
P = Position

Boxplot for Outliers Detection

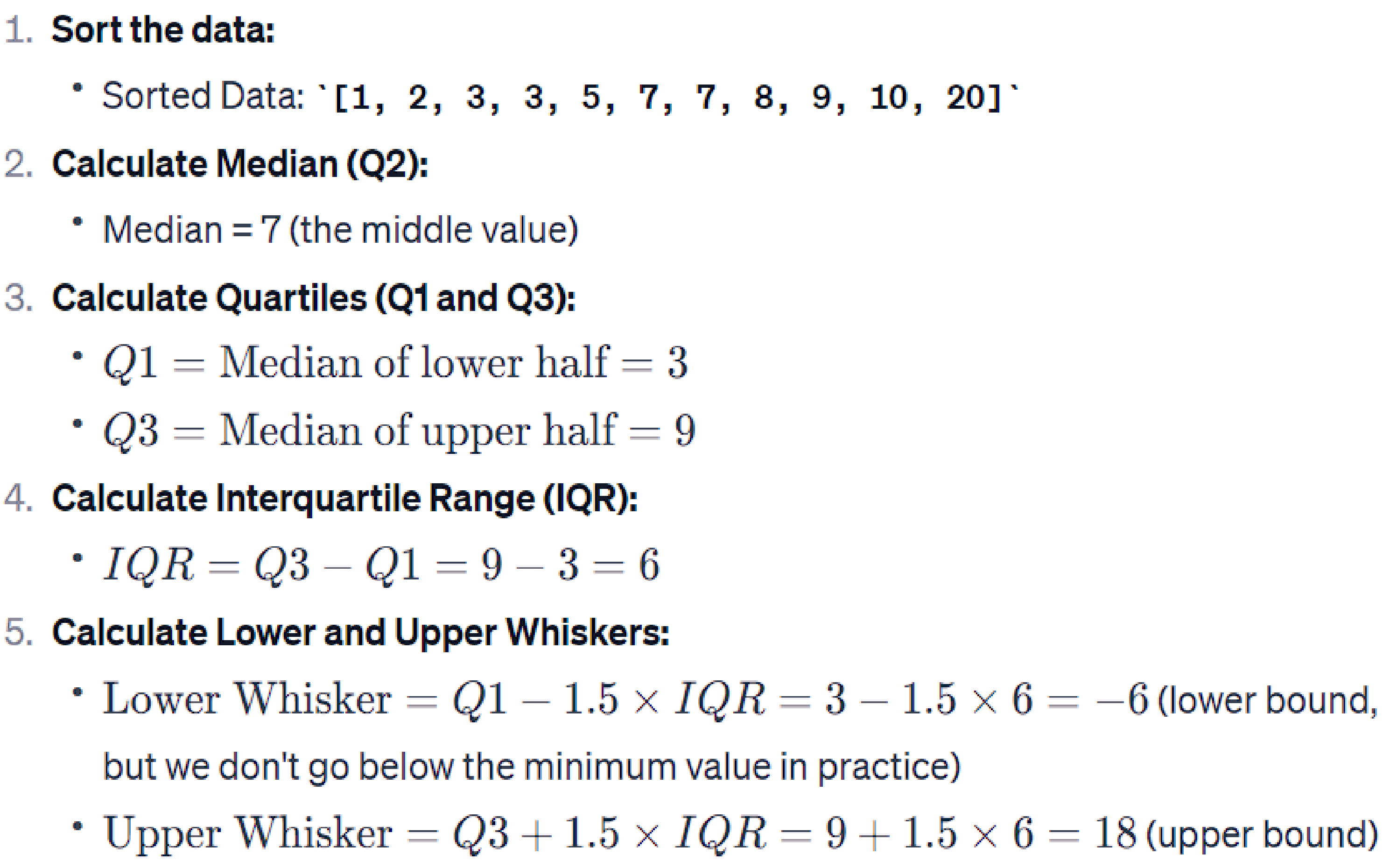
Construct Boxplot:

Data = [1,2,3,3,5,7,7,8,9,10,20]

- Min = 1
- Q1 = (25/100)*(11+1) = 3rd value = 3
- Median = (50/100)*(11+1) = 6th value = 7
- Q3 = (75/100)*(11+1) = 9th value = 9
- IQR = 9-3 = 6
- Max = 20

- Lower Whisker = Q1 – (1.5*IQR)
  - = 3 – 9
  - = - 6

- Upper Whisker = Q3 + (1.5*IQR)
  - = 9 + 9
  - = 18



1. **Sort the data:**
   - Sorted Data: `[1, 2, 3, 3, 5, 7, 7, 8, 9, 10, 20]`
2. **Calculate Median (Q2):**
   - Median = 7 (the middle value)
3. **Calculate Quartiles (Q1 and Q3):**
   - $Q1 = \text{Median of lower half} = 3$
   - $Q3 = \text{Median of upper half} = 9$
4. **Calculate Interquartile Range (IQR):**
   - $IQR = Q3 - Q1 = 9 - 3 = 6$
5. **Calculate Lower and Upper Whiskers:**
   - $\text{Lower Whisker} = Q1 - 1.5 \times IQR = 3 - 1.5 \times 6 = -6$ (lower bound, but we don't go below the minimum value in practice)
   - $\text{Upper Whisker} = Q3 + 1.5 \times IQR = 9 + 1.5 \times 6 = 18$ (upper bound)

Boxplot for Outliers Detection

To create a boxplot, you can follow these steps:

1. Gather your dataset: Collect the numerical data that you want to visualize using a box plot.

2. Determine the key components: Identify the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum values of your dataset.

3. Calculate the interquartile range (IQR): Subtract Q1 from Q3 to obtain the IQR.

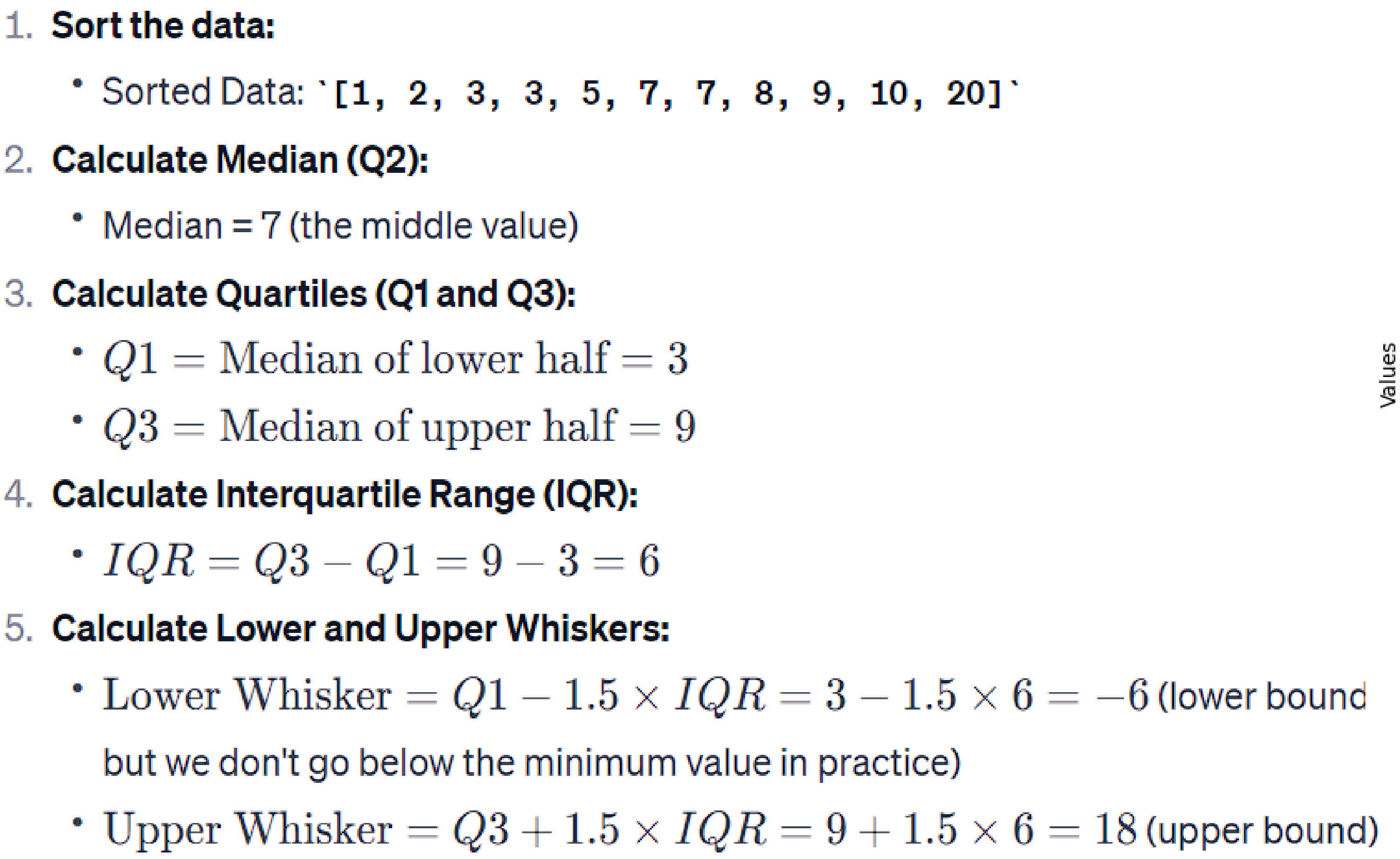
4. Identify any outliers: Determine if any values fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR. These values are considered outliers and may be plotted individually as points.

5. Set up the box plot: Draw a number line or axis to represent the range of your dataset. Place a box that spans from Q1 to Q3 on the number line. Draw a line within the box to represent the median (Q2).

6. Add whiskers: Extend lines, known as "whiskers," from the box to the minimum and maximum values that are not considered outliers.

7. Plot outliers: If there are any outliers, plot them individually as points outside the whiskers.

Boxplot for Outliers Detection



1. **Sort the data:**
   - Sorted Data: `` `[1, 2, 3, 3, 5, 7, 7, 8, 9, 10, 20]` ``
2. **Calculate Median (Q2):**
   - Median = 7 (the middle value)
3. **Calculate Quartiles (Q1 and Q3):**
   - $Q1 = \text{Median of lower half} = 3$
   - $Q3 = \text{Median of upper half} = 9$
4. **Calculate Interquartile Range (IQR):**
   - $IQR = Q3 - Q1 = 9 - 3 = 6$
5. **Calculate Lower and Upper Whiskers:**
   - $\text{Lower Whisker} = Q1 - 1.5 \times IQR = 3 - 1.5 \times 6 = -6$ (lower bound but we don't go below the minimum value in practice)
   - $\text{Upper Whisker} = Q3 + 1.5 \times IQR = 9 + 1.5 \times 6 = 18$ (upper bound)

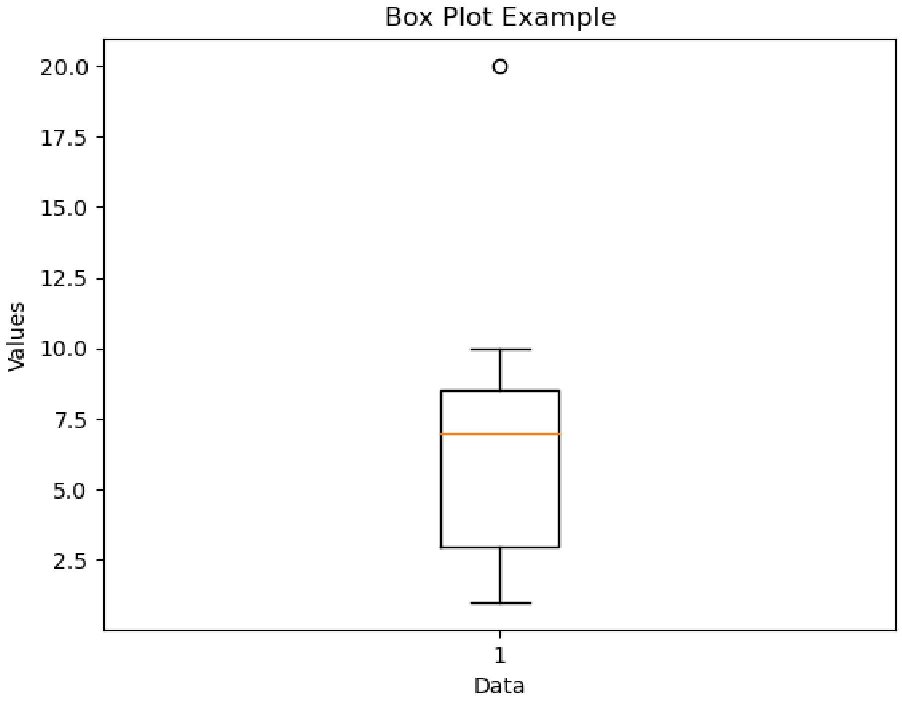


Fig: Boxplot

Good Luck!