

Statistics Foundation

- Statistics and its role in Data Science
- Statistics and its Types
 - Descriptive Statistics
 - Inferential Statistics
- Population And Sample
- Sampling Techniques
- Descriptive Statistics
 - Graphical Representation
 - Frequency Distribution
 - Bar Graph, Histograms
 - Measures of Central Tendency
 - Mean, Median, Mode
 - Measure of Dispersion/Spread
 - Variance, Standard Deviation
 - Percentiles and Quartiles
 - Range, Interquartile Range
 - Box and Whisker plot for Outlier Removal
 - Five number summary
- Anomaly v/s Outlier

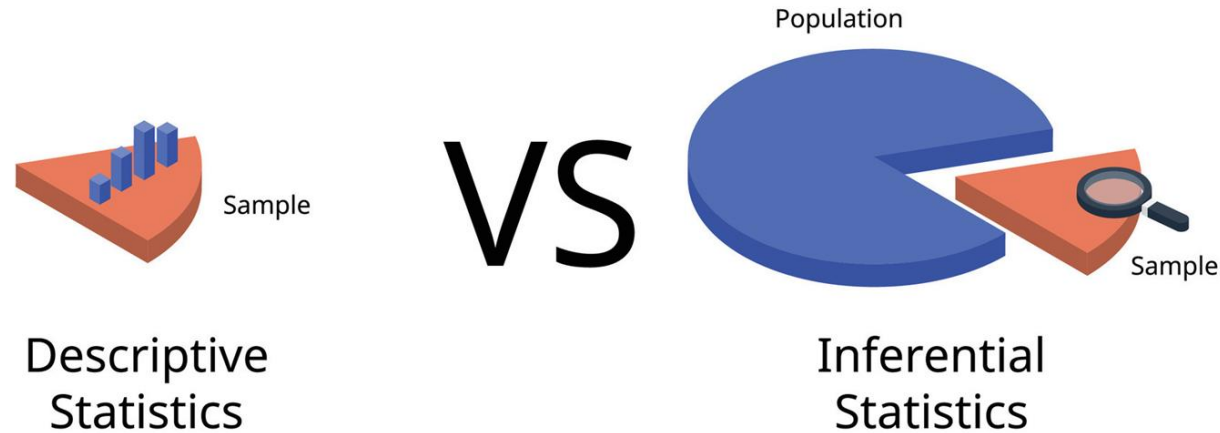
Statistics Advance

- Inferential Stats
- Z-Score (Standard Score)
- Transformation
- Standardization
- Normalization (Min Max Scaling)
- Log Transformation
- Sampling Distribution
- Standard Error
- Estimate and Estimator
- Properties of Good Estimator
 - Confidence interval
- Q-Q Plot
- Central Limit Theorem
- Hypothesis testing
 - Significance Value
 - p-value
 - Type 1 and Type 2 error
- Statistical Tests

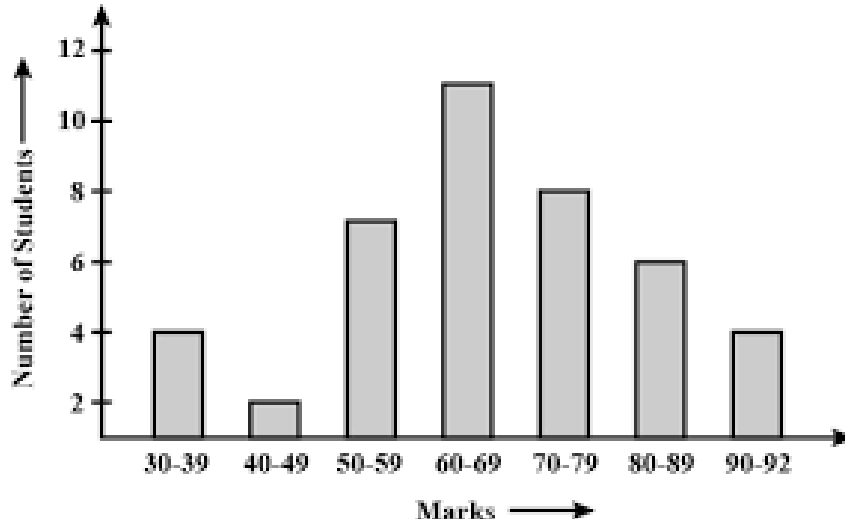

Statistics and Its role in Data Science

- Statistics is the branch of mathematics that involves collecting, analyzing, interpreting, presenting, and organizing data.
 - It provides tools and methods to make sense of data, allowing us to understand patterns, make decisions, and draw conclusions based on numerical information.
1. Data Collection and Sampling
 2. Data Cleaning and Preprocessing
 3. Data Summarization and Exploration
 4. Inferential Statistics
 5. Modeling and Prediction (Machine Learning)
 6. Decision Making (A/B Testing)

Descriptive Statistics vs Inferential Statistics



Aspect	Descriptive Statistics	Inferential Statistics
Objective	Summarize data to highlight its main features clearly and understandably, focusing solely on characteristics within the sample.	Use sample data to make predictions and draw conclusions about a larger population. Test hypotheses and make informed decisions about population characteristics based on this sample information.

Aspect	Descriptive Statistics	Inferential Statistics																
Methods	Measures of central tendency, dispersion, and graphical representations.	Hypothesis testing, confidence intervals, regression analysis, etc.																
Example	 <table border="1"><thead><tr><th>Marks</th><th>Number of Students</th></tr></thead><tbody><tr><td>30-39</td><td>4</td></tr><tr><td>40-49</td><td>2</td></tr><tr><td>50-59</td><td>7</td></tr><tr><td>60-69</td><td>11</td></tr><tr><td>70-79</td><td>8</td></tr><tr><td>80-89</td><td>6</td></tr><tr><td>90-92</td><td>4</td></tr></tbody></table>	Marks	Number of Students	30-39	4	40-49	2	50-59	7	60-69	11	70-79	8	80-89	6	90-92	4	
Marks	Number of Students																	
30-39	4																	
40-49	2																	
50-59	7																	
60-69	11																	
70-79	8																	
80-89	6																	
90-92	4																	
Other Examples	Calculate mean income, create histograms, bar charts.	Test if a new drug is effective, estimate population parameters, predict future outcomes.																

Example

Suppose you have a sample of heights from a school, denoted as:

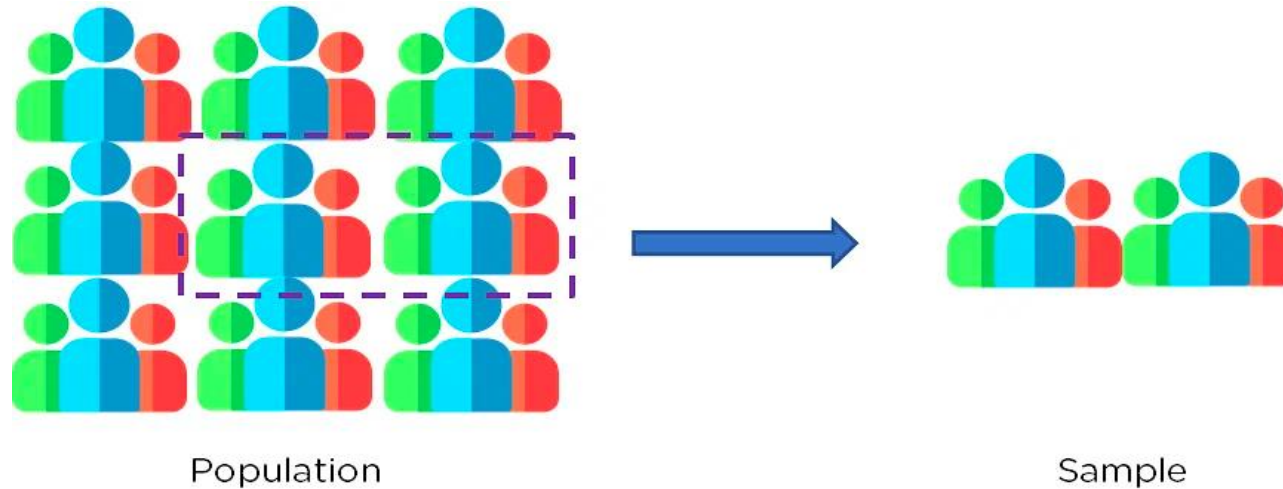
$$X=\{x_1, x_2, \dots, x_{50}\}$$

Descriptive Statistics: Mean (Average), Histogram

Inferential Statistics:

- You calculate a confidence interval for the mean height, which gives you a range of heights within which you can be confident the true population mean falls.
- You perform a hypothesis test to determine whether the average height of students in your school is different from the national average.
- Inferential statistics can also help you decide how large your sample should be to make reliable inferences about the entire student population.
- Answering questions like “Are marks of students of this classroom similar to marks of Maths classroom in the college?”

Population (N) And Sample (n)

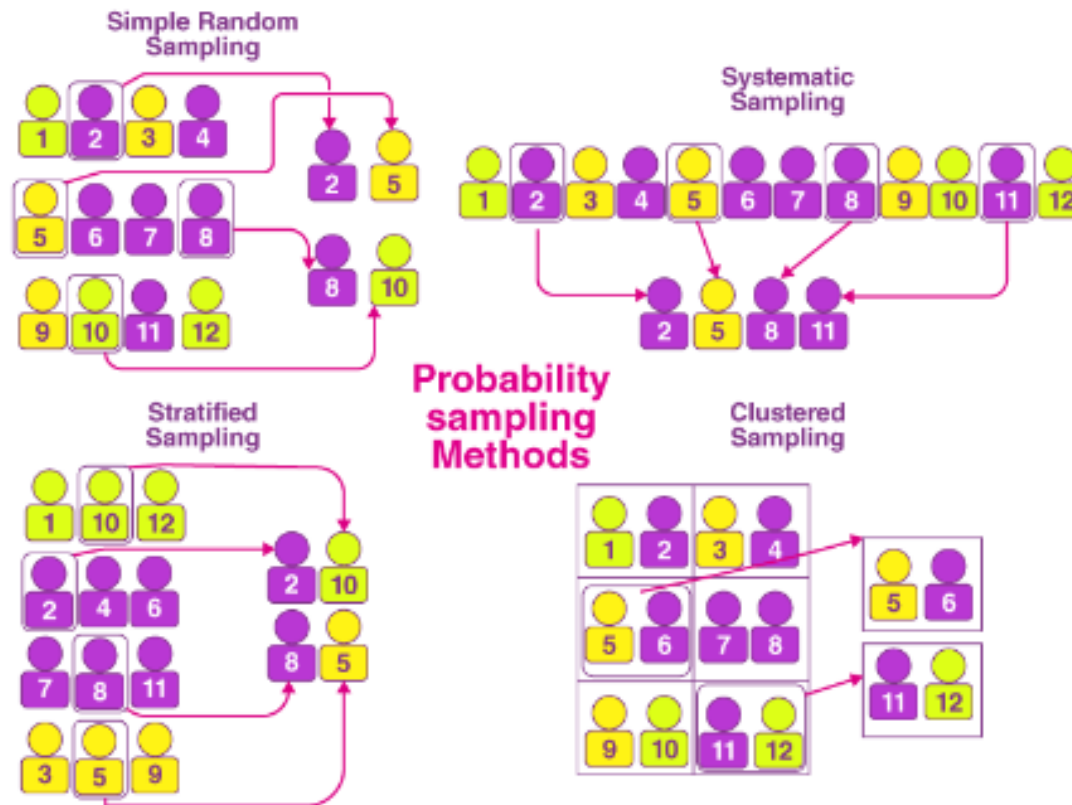


Population: The entire group of individuals, objects, or data points under study. Example: Studying average income among all professionals in a city.

Sample: A representative subset of the population used for drawing conclusions. Example: Randomly selecting 500 professionals to estimate city-wide income averages.

Sampling Techniques

The choice of sampling method depends on the research objectives, available resources, and the desired level of representation within the sample. The key to effective sampling is ensuring that the sample is representative and unbiased, allowing for valid generalizations to be made about the population.



Each of these sampling techniques has its own advantages, limitations, and appropriate use cases.

Simple Random Sample	Systematic Sampling	Stratified Sampling	Cluster Sampling
Each individual in the population has an equal chance of being selected.	Sampling is performed by selecting every kth element from a list.	The population is divided into subgroups (strata), and random samples are drawn from each stratum.	Population divided into clusters by criteria; random clusters chosen, all individuals within chosen clusters sampled.

Randomly selecting students from a class list to conduct a survey.	Simple Random Sampling (SRS)
Dividing a population of employees into different job roles (strata) and selecting random samples from each stratum to ensure representation of gender in all job roles.	Stratified Sampling
Selecting every 10th customer from a list of customers in a database for a customer satisfaction survey.	Systematic Sampling
Survey reading habits by randomly selecting 5 out of 20 schools in a district and surveying all students in the chosen schools.	Cluster Sampling

Frequency Distribution

Continuous Frequency Distribution

Ex 1 : Here are the exam scores for a class of 30 students. The exam scores range from 60 to 100.

80, 85, 72, 90, 92, 78, 85, 88, 70, 78, 88, 92, 95, 80, 75,
85, 92, 88, 80, 78, 75, 90, 92, 80, 85, 70, 78, 88, 85, 75

Score Range	Frequency
60-70	0
70-80	10
80-90	13
90-100	7

Discrete Frequency Distribution

Ex 2 : List of favourite colors of students in a class:

Red,Blue,Green,Red,Blue,Green,Green,Yellow,Blue,
Purple

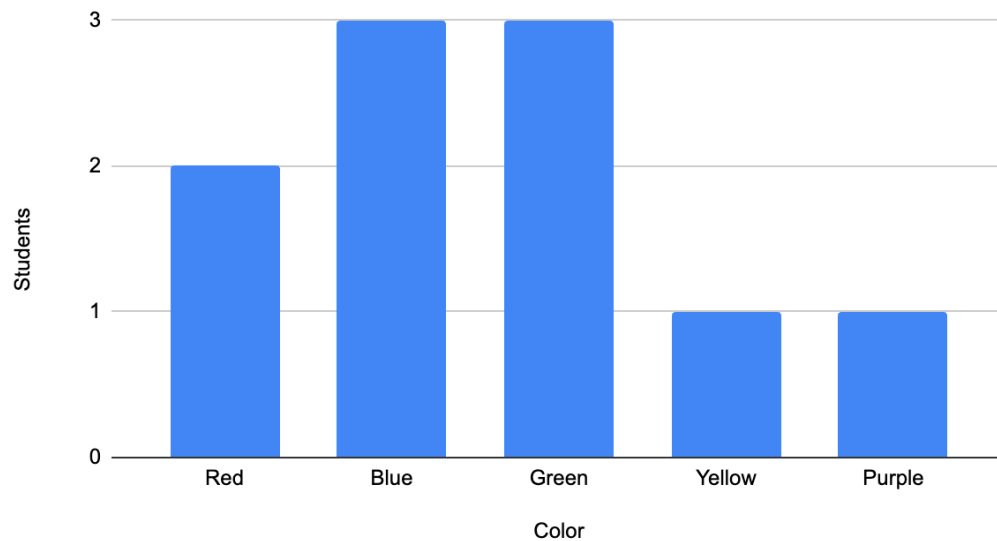
Color	Frequency
Red	2
Blue	3
Green	3
Yellow	1
Purple	1

Bar Graphs and Histograms

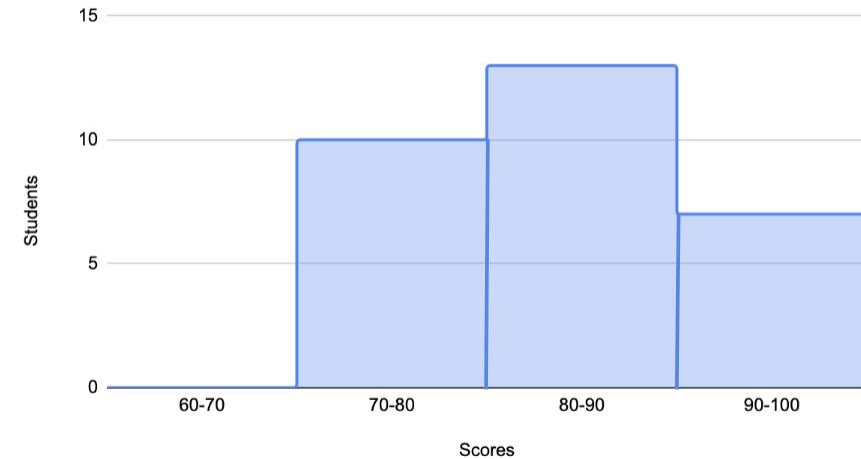
Bar Graph: Represents categorical or discrete data using bars of equal width, where the height of each bar corresponds to the frequency or count of each category.

Histogram: Visualizes the distribution of continuous data by grouping values into intervals (bins). The bars in a histogram touch each other to show that the data is continuous, and the height of each bar represents the frequency of values within that interval.

Students vs Color BarGraph



Students vs Score Histogram



Use Cases of Stats

1. In a survey, we collected the ages of participants. What is the average age of the participants?
2. In a marathon race, what is the "typical" finishing time for runners?
3. In a survey, we collected students placement data. Which company is the most popular among the students?
4. In a college, how do we get to know average placement range of students?
5. We have test scores from two different schools. How do we compare the spread of scores between the two schools?
6. In a sales dataset, what is the difference between the highest and lowest sales figures in a given period?
7. We have employee salaries for a company. How can we understand the distribution of salaries within the company?
8. In a test score dataset, how does a student's score compare to the rest of the students? Is it in the top 10%?

Measures of Central Tendency

A measure of central tendency identifies the center of a data set, summarizing its central position.

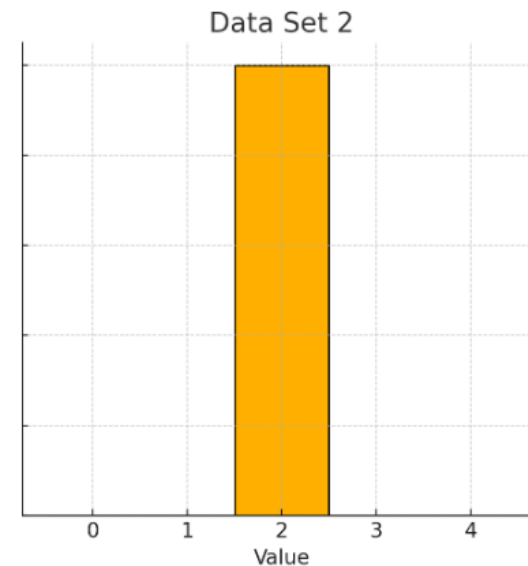
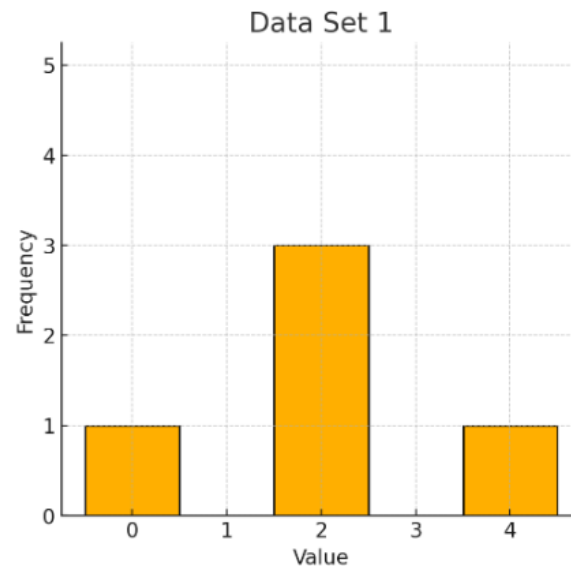
Measures	Mean/Average	Median	Mode
Definition	Sum of all values divided by the number of observations.	Middle value when dataset is arranged.	Most Frequent Value
Calculation	$\mu = (x_1 + x_2 + \dots + x_n) / n$	<p>n is odd, median = value at $(n+1)/2$;</p> <p>n is even, median = average of values at $n/2$ and $(n/2)+1$</p>	<p>The mode may not be unique or may not exist in some datasets.</p> <p>There can be unimodal, multimodal or no mode models.</p>
Usefulness	Appropriate for a roughly symmetrical distribution, such as the normal distribution.	Appropriate measure for skewed data.	Useful when looking for the most common value or generally when the data is categorical.
Example	<p>Mean of height of students</p> <p>5,5.5,4,4.5,5.2,5.8</p> <p>Mean = 5</p>	<p>Median of annual salaries in college placement where 100L is an outlier:</p> <p>5L,6L,10L,3L,8L,7L,100L</p> <p>Median = 7L</p>	<p>M, M, M, F, M, F,</p> <p>Mode = M</p>

Why Measures of Central Tendency not sufficient Enough ?

Central tendency measures (mean, median, mode) can be identical in different distributions but don't reveal data spread or scatter.

Dataset 1 $\rightarrow \{0, 2, 2, 2, 4\}$ \rightarrow Mean = 2, Median = 2, Mode = 2

Dataset 2 $\rightarrow \{2, 2, 2, 2, 2\}$ \rightarrow Mean = 2, Median = 2, Mode = 2



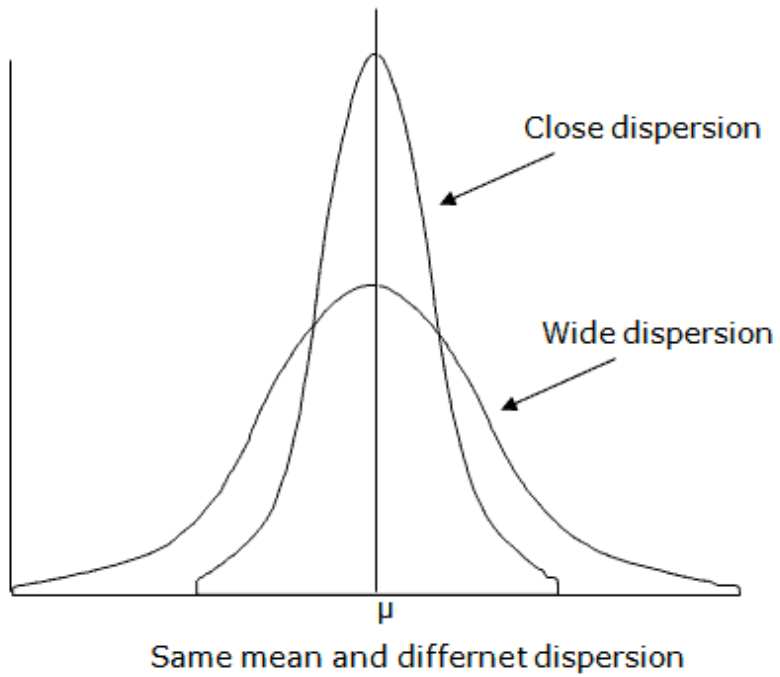
And Hence there was a need for Measures of Dispersion (to measure the spread) !

Measures of Dispersion (Spread)

Measures	Variance	Standard Deviation
	Essential statistical tools for quantifying and understanding the spread of data.	
Definition	Variance is the mean of the squared differences from the mean.	Standard deviation is the square root of the variance understanding how far a value is from the mean.
Formula	<p>Population Variance</p> $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$ <p>Sample Variance</p> $s^2 = \frac{\sum (x_i - \bar{x})^2}{(n - 1)}$	<p>Population Standard Deviation</p> $\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$ <p>Sample Standard Deviation</p> $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n - 1)}}$

Why square of difference from mean in Variance calculation?

We could have also taken an absolute value of differences from mean to avoid cancelling of -ve and +ve deviation but the use of squared differences has some mathematical advantages, such as simplifying calculations and being friendly to certain statistical tests. Also Differentiation of modulus is not defined at 0.



More Variance -> More spread -> More dispersion

Why n-1 in Sample Variance and Sample Standard Deviation?

Sample Variance

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

Sample Standard Deviation

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

- When working with a sample instead of the entire population, we use sample statistics to estimate population parameters.
- Bessel's correction adjusts for the tendency of sample data to underestimate population variability.
- It involves dividing by n-1 instead of n in the denominator of the formula.
- This adjustment provides a more accurate estimation of population variability by compensating for the smaller size of the sample compared to the entire population.

Example

Consider the following dataset representing the test scores of a group of students:
78, 85, 92, 88, 75, 80, 92, 80, 85, 70

Mean = 80.5

Variance ≈ 56.528

Standard Deviation $\approx \sqrt{56.528} \approx 7.52$

The "**1st std**" represents the range of values within one standard deviation of the mean. In our example, it would include values within 1 standard deviation of 80.5, so the range is approximately $[80.5+7.52, 80.5-7.52]$

The "**2nd std**" represents the range of values within two standard deviations of the mean. In our example, it would include values within 2 standard deviations of 80.5, so the range is approximately $[80.5+2(7.52), 80.5-2(7.52)]$

The "**3rd std**" represents the range of values within three standard deviations of the mean, and so on. ■

Try to Answer !

1. Why use Standard Deviation when we have Variance?
2. Is Standard deviation robust to outliers?

Measures of Dispersion (Spread)

	Percentile	Quartile																					
Definition	<p>Percentile divides data into 100 equal parts.</p> <p>Percentiles are a way to express the relative position of a value within a dataset.</p> <p>Also known as a centile, it indicates the percentage of values in a dataset that fall below a specific value.</p>	<p>Quartiles divide data into four equal parts.</p> <div><div>1st half</div><div>2nd half</div><table><tr><td>24</td><td>28</td><td>31</td><td>32</td><td>36</td><td>45</td><td>54</td><td>55</td><td>58</td><td>60</td><td>63</td></tr><tr><td colspan="2">minimum</td><td colspan="2">lower quartile Q1</td><td colspan="2">median Q2</td><td colspan="2">upper quartile Q3</td><td colspan="2">maximum</td></tr></table></div>	24	28	31	32	36	45	54	55	58	60	63	minimum		lower quartile Q1		median Q2		upper quartile Q3		maximum	
24	28	31	32	36	45	54	55	58	60	63													
minimum		lower quartile Q1		median Q2		upper quartile Q3		maximum															
Formula	<p>1. Order the dataset with total N data points in ascending order.</p> <p>2. Calculate the rank (position) of the desired percentile(P) using the formula:</p> <p>Rank = (Percentile / 100) * (N + 1)</p> <p>3. Locate the data point at the calculated position (rounded if necessary). This is the percentile value.</p>	<p>Q1: 25th percentile</p> <p>Q2 (the median): 50th percentile</p> <p>Q3: 75th percentile</p>																					

Example

Marks of 10 students : 45 55 60 68 70 75 78 82 90 92

Find percentile of students with marks = 60

Percentile = $(3/10) * 100 = 30^{\text{th}}$ i.e. 30%ile.

What value exists at percentile ranking of 25th?

index = $(25/100) * 11 = 2.75$

Since the position (2.75) falls between the 2nd (55) and 3rd (60) data point:

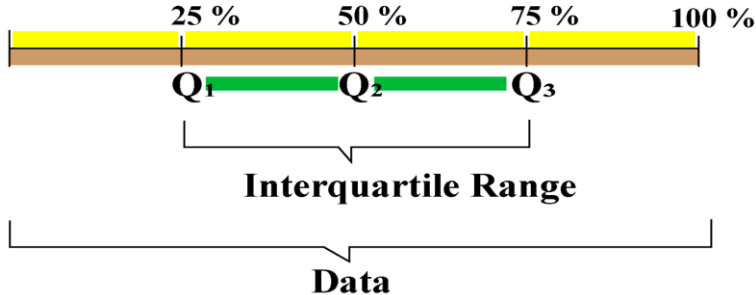
Interpolation (if used): Statistical software might use interpolation to account for the non-integer position. This could lead to a value between 55 and 60.

Nearest neighbor (if used): Some methods might simply take the closest data point. In this case, the 25th percentile would be 55.

Without knowing the specific method used, we can't definitively say whether the 25th percentile is exactly 55 or a value between 55 and 60. However, given the small dataset and closeness of the values, 55 is a reasonable estimate for the 25th percentile.

Note : The method used for calculation might vary depending on the data and available tools.

Measures of Dispersion (Spread)

	Range	Interquartile Range
Definition	<p>The range is the difference between the maximum and minimum values, showing the dataset's overall spread but not the central spread.</p>	<p>The IQR is a measure of statistical dispersion that quantifies the spread of the middle 50% of data.</p> <p>$IQR = Q3 - Q1$</p> <p>It is useful for identifying the variability within the central portion of the dataset.</p>
Formula	<p>16, 24, 22, 25, 26, 27, 28, 23</p> <p>Range = max - min</p> <p>Range = 28 - 16 = 12</p>	

Try to Answer !

1. Can percentile be exactly 100?

No, in the strictest sense, a percentile cannot be exactly 100. Here's why:

- **Percentile Definition:** A percentile represents the value below which a specific percentage (e.g., 25th, 50th, 75th) of the data falls.
- **100th Percentile Interpretation:** The 100th percentile would indicate the value below which 100% of the data falls.

Highest Value: In practice, the highest value in the dataset is often referred to as being at the "99th percentile" or close to it. This implies that a very high percentage (but not necessarily 100%) of the data falls below it.

Rounding: Sometimes, percentile calculations might involve rounding, especially for non-integer positions. In such cases, a data point at the very top could be rounded up to the 100th percentile. However, this is a rounding convenience and doesn't strictly adhere to the definition of a percentile.

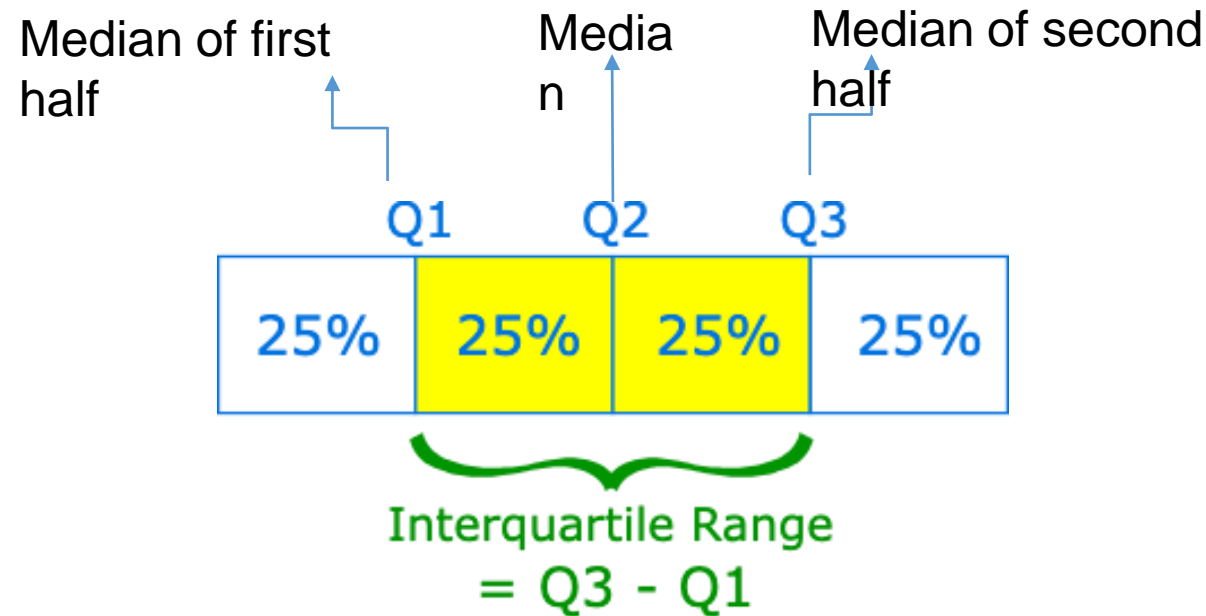
Different Methods of Quartile Calculation

Scores : 72,85,68,92,78,60,88,75,96,45

Step 1: Sort the dataset in ascending order : 45,60,68,72,75,78,85,88,92,96

Method 1: Calculate Quartiles using median of medians.

Method 2: Calculate Quartiles using percentile/quantile method.



Method : Calculate Quartiles using percentile/quantile method.

Score of 10 students : 72,85,68,92,78,60,88,75,96,45

Sort the dataset in ascending order : 45,60,68,72,75,78,85,88,92,96

Step 1: Calculate **Quartiles**

Calculate the first quartile (Q1):

- $Q1 = (25/100) * (10+1) = 2.75\text{th Rank} \rightarrow \text{Between the 2nd and 3rd data points} = (60 + 68)/2 = 64$

Calculate the second quartile (Q2):

- $Q2 = (50\text{th percentile}) = \text{median}$
- $Q2 = (50/100) * (10+1) = 5.5\text{th Rank} \rightarrow \text{Between 5th and 6th data point} = (75+78) / 2 = 76.5$

Calculate the third quartile (Q3):

- $Q3 = (75/100) * (10+1) = 8.25\text{th Rank} \rightarrow \text{Between the 8th and 9th data points} = (88 + 92)/2 = 89$

Step 2: Calculate the **IQR** = $Q3 - Q1 = 89 - 64 = 25$

Step 3: Calculate **Range** = $96 - 45 = 51$

Practice

Mean: What is the average age of survey participants?

Median: What is the "typical" marathon finishing time? Use the fastest or slowest time?

Mode: Which company is most popular among students in a placement survey?

Median: How do we find the average placement range in a college?

IQR/Variance: How do we compare the spread of test scores between two schools?

Range: What is the difference between the highest and lowest sales figures in a sales dataset?

Quartiles: How can we understand the salary distribution in a company?

Percentiles: How does a student's test score compare to the rest? Is it in the top 10%?

Mode: Does a mode exist in this data (1,1,2,2,3,3,4,4)? Multiple modes exist (all are modes).

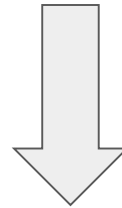
Sensitivity to Outliers

Measure	Sensitiveness	Reason
Mean	Yes	Outliers can significantly affect the average value by skewing the data distribution.
Median	No	The median is the middle value and is not influenced by extreme values.
Mode	No	The mode is the most frequent value and is not affected by the magnitude of outliers.
Variance and Std	Yes	These measures depend on the squared differences from the mean, so outliers can increase them significantly.
Percentiles	No	Percentiles divide the data into equal parts and are not affected by extreme values.
Range	Yes	The range is the difference between the maximum and minimum values, so outliers directly affect it.
IQR	No	The IQR measures the spread of the middle 50% of the data, excluding outliers.

Are Measures of Dispersion sufficient Enough ?

Dataset with Same
Central Tendency

Dataset1 -> 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8, 8, 9
 Dataset2 -> 1, 2, 3, 4, 5, 5, 6, 7, 8, 9
 Mean = Median = Mode = 5 for both Dataset
 Var (Dataset1) = 4.0 and Var (Dataset2) = 6.0



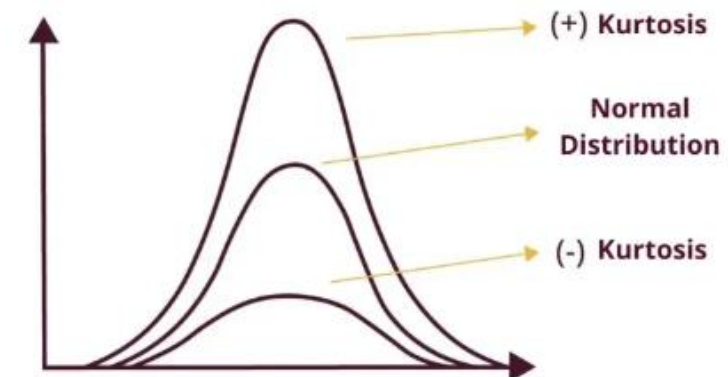
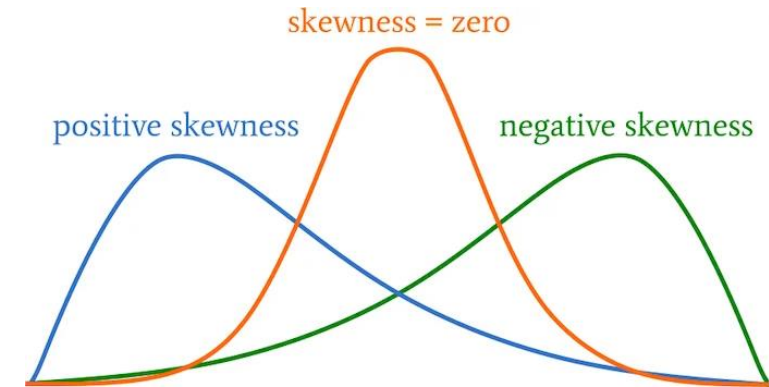
Dataset with Same Central
Tendency and Dispersion

Dataset1 -> 0.1 , 1.33, 1.33, 2.55, 2.55, 2.55, 3.78, 3.78, 3.78, 3.78, 5.0, 5.0, 5.0,
 5.0, 5.0, 6.22, 6.22, 6.22, 6.22, 7.45, 7.45, 7.45, 8.67, 8.67, 9.9
 Dataset2 -> 1, 2, 3, 4, 5, 5, 6, 7, 8, 9
 Mean = Median = Mode = 5 for both Dataset
 Var (Dataset1) = 6.0 and Var (Dataset2) = 6.0

And Hence there was a need for Skewness and Kurtosis

Quick Comparison

Statistic	Definition
Central Tendency (Degree 1)	Central tendency measures the central point of a dataset.
Variance (Degree 2)	Measures of dispersion in statistics capture how data is spread out around central values like the mean or median.
Skewness (Degree 3)	Skewness measures the asymmetry of the distribution. It shows if the data is more spread out on one side of the mean than the other.
Kurtosis (Degree 4)	Kurtosis measures the "tailedness" of the distribution, or how heavy or light the tails of the distribution are compared to a normal distribution. High kurtosis(heavy tail) means more outliers and a sharper peak, while low kurtosis(light tail) means fewer outliers and a flatter peak.



Skewness and Kurtosis

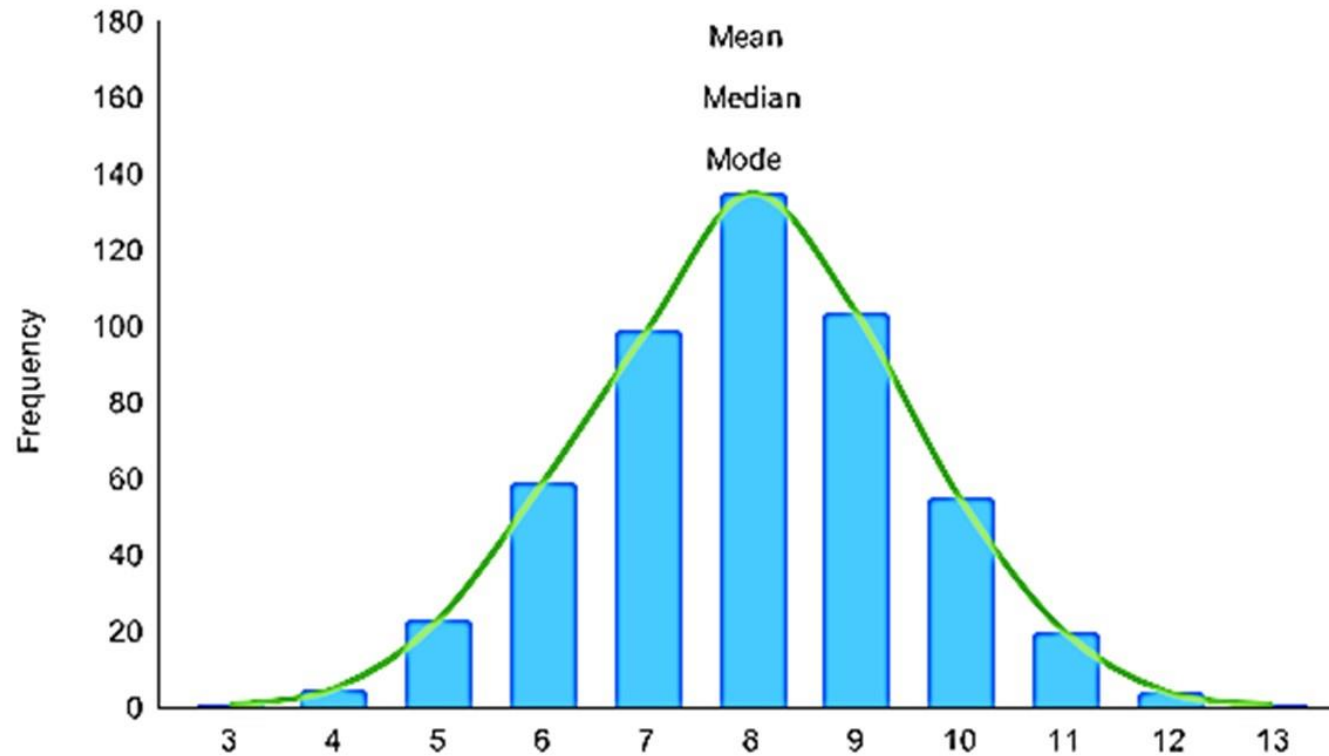
$$\text{Skew} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

$$\text{Kurtosis} = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Where:

- n is the number of observations in the sample.
- x_i are the individual data points.
- \bar{x} is the sample mean.
- s is the sample standard deviation.

Interview Question



Symmetric Distribution

Example : Heights of adult men

Mean = Median = Mode

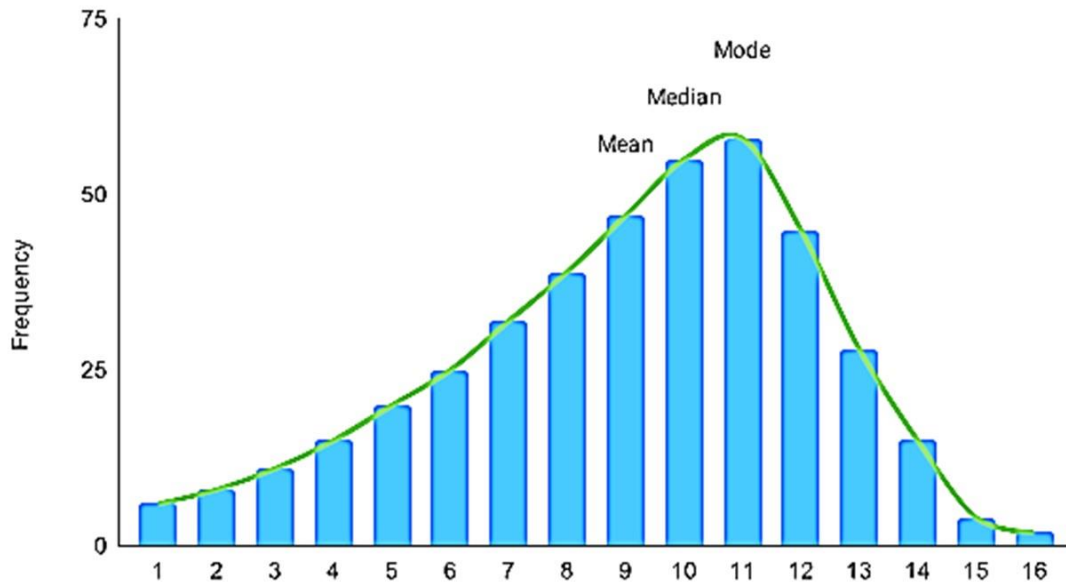
In such a case, we emphasize the mean value of the distribution.

Left Skewed

Example : Age of death from natural causes.

Mean < Median < Mode

In such a case, we emphasize the median value of the distribution.

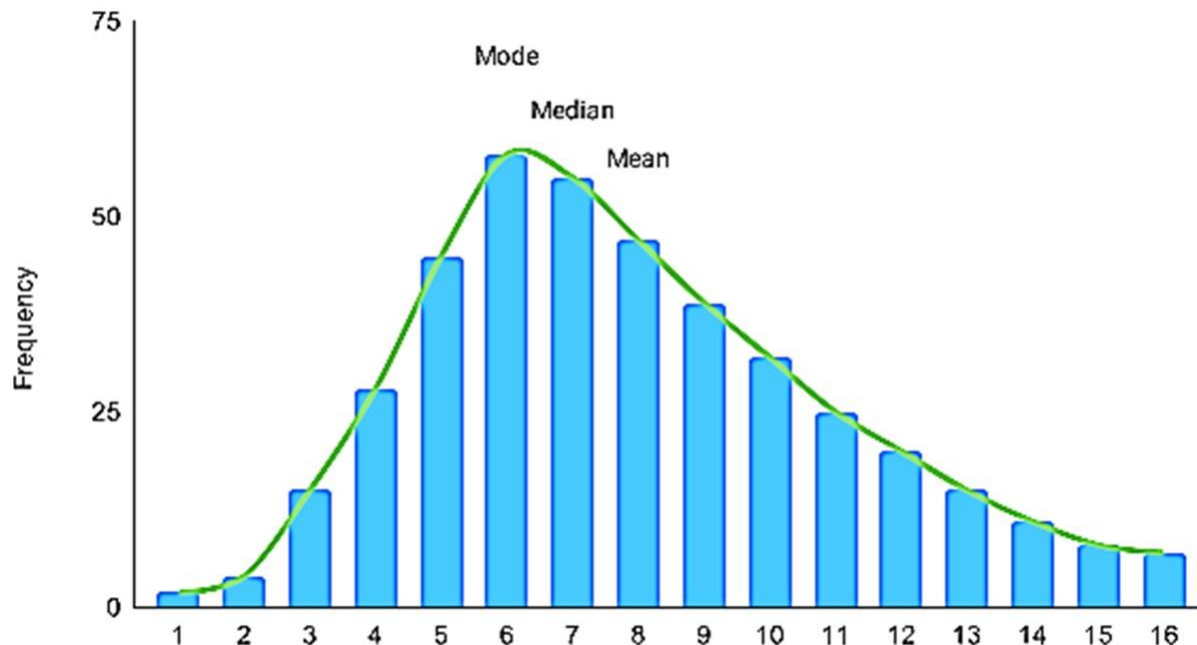


Right Skewed

Example : Salaries of employees, where higher-earners provide a false representation of the typical income if expressed as mean.

Mean > Median > Mode.

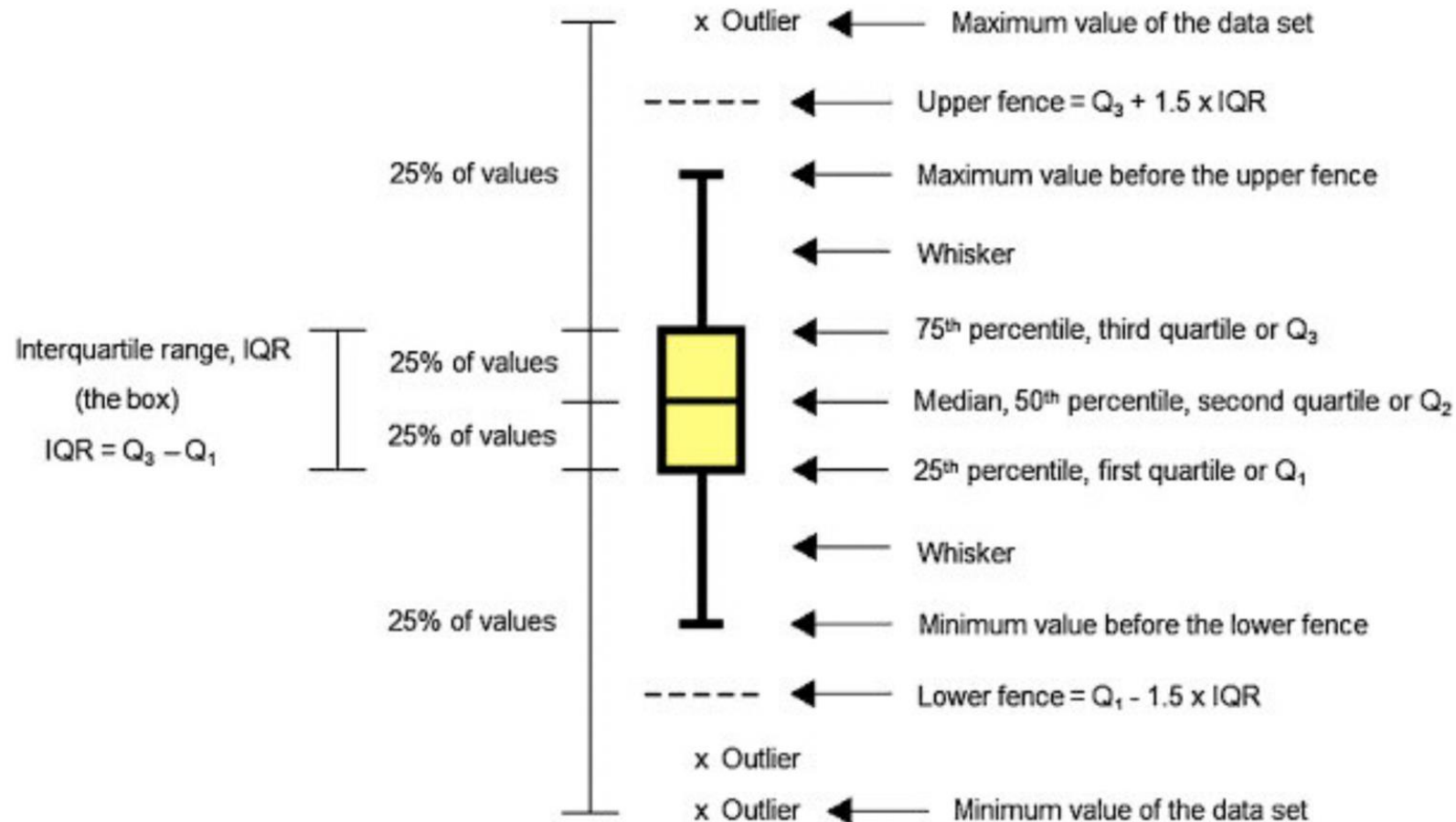
In such a case, we emphasize the median value of the distribution.



Box and Whisker Plot (Five number summary)

A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. They provide valuable insights into central tendency, variability, and the presence of outliers.

minimum, first quartile, median, third quartile, and maximum.



Detecting Outliers and Data Visualisation using BoxPlot

1,2,2,2,3 |,3,4,5,5,5 |,6,6,6,6,7,| 8,8,9,11,27

$Q1 = .25 \times (21) = 5.25$ th index = 3

$Q2 = \text{median} = (5+6)/2 = 5.5$

$Q3 = .75 \times (21) = 15.75$ th index = 7.75

$IQR = Q3 - Q1 = 4.75$

Lower Fence = $3 - 1.5 \times (4.75) = -4.125$

Upper Fence = $7.75 + 1.5 \times (4.75) = 14.875$

Hence anything outside $[-4.125, 14.875]$ will be considered an outlier.

Minimum : 1

Q1 : 3.0

Median : 5.5

Q3 : 7.75

Maximum : 27



Example

Example: Finding the five-number summary

[-10], [2], [4], [3], [8], [7], [9], [15], [11], [-5]

Make a box plot of the data.

Step 1: Order the data from smallest to largest.

-10, -5, 2, 3, 4, 7, 8, 9, 11, 15

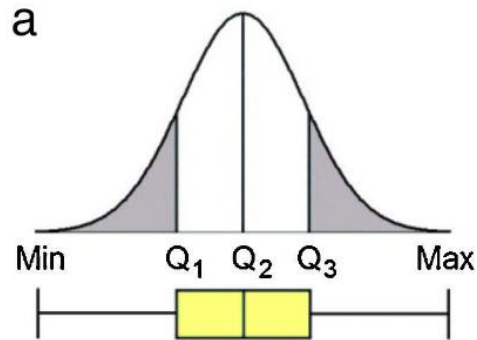
Step 2: Find the median.

The median is $(4+7)/2 = 5.5$

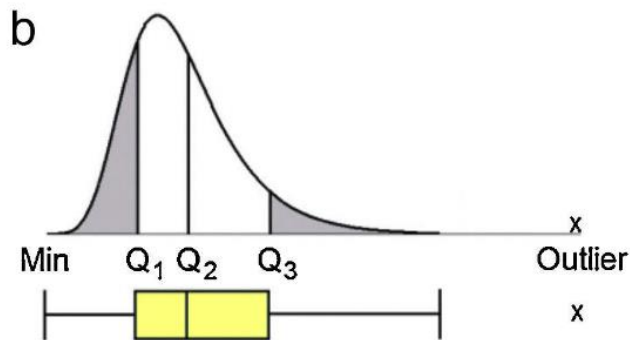
Step 3: Find the quartiles.

Why 1.5 IQR?

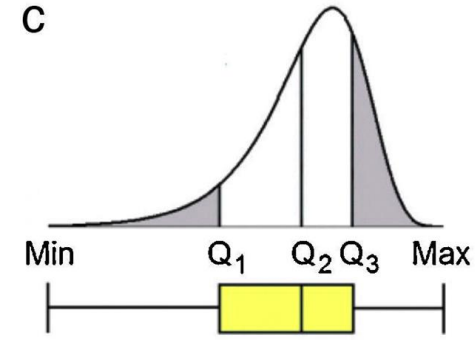
- The 1.5 IQR rule is effective for identifying outliers in both symmetrical and asymmetrical distributions.
- The IQR represents a robust measure of spread, accounting for the central bulk of the data.
- **Sensitivity to Outliers:** Using $1.5 * \text{IQR}$ strikes a balance between detecting outliers that are potentially extreme while still being reasonably sensitive to slight variations in the data distribution.
- The choice of multiplier (1.5 or 3.0) can be adjusted based on the context and the specific characteristics of your data like dealing with data that is known to have extreme values or if a different level of sensitivity to outliers is desired.



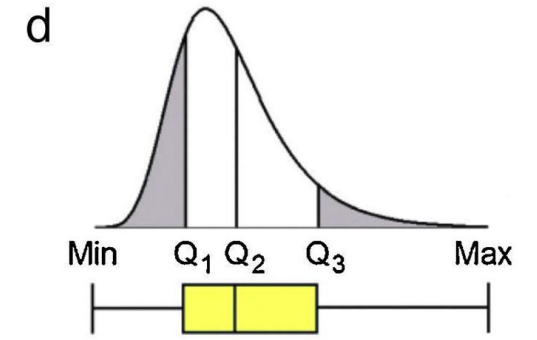
Symmetric



Asymmetric
(positive or right skewed)



Asymmetric
(negative or left skewed)



Asymmetric
(positive or right skewed)

Anomaly vs Outlier

Aspect	Anomalies	Outliers
Definition	A data point or event that is unexpected, surprising, or abnormal based on the context or historical data.	Data points that significantly differ from the other observations in a dataset.
Detection	Identified using machine learning, clustering, or domain-specific rules.	Identified using statistical methods such as Z-scores, interquartile range (IQR), or box plots.
Implications	More relevant in fields like finance, network security, and health monitoring.	Depending on context, they may be removed, corrected, or retained.
Example	A company's stock price, typically stable at \$100 per share, suddenly drops to \$50 in a single trading day without evident cause, marking an anomaly.	Consider a dataset representing the daily number of website visitors over a month: 100, 120, 110, 105, 115, 125, 130, 100, 110, 115, 105, 120, 125, 130, 100, 110, 115, 105, 120, 125, 130, 100, 110, 115, 105, 120, 125, 130, 1000

Outliers denote statistical deviations within dataset, while anomalies imply unexpected or significant deviations beyond statistical norms, with context playing a crucial role in distinguishing them.