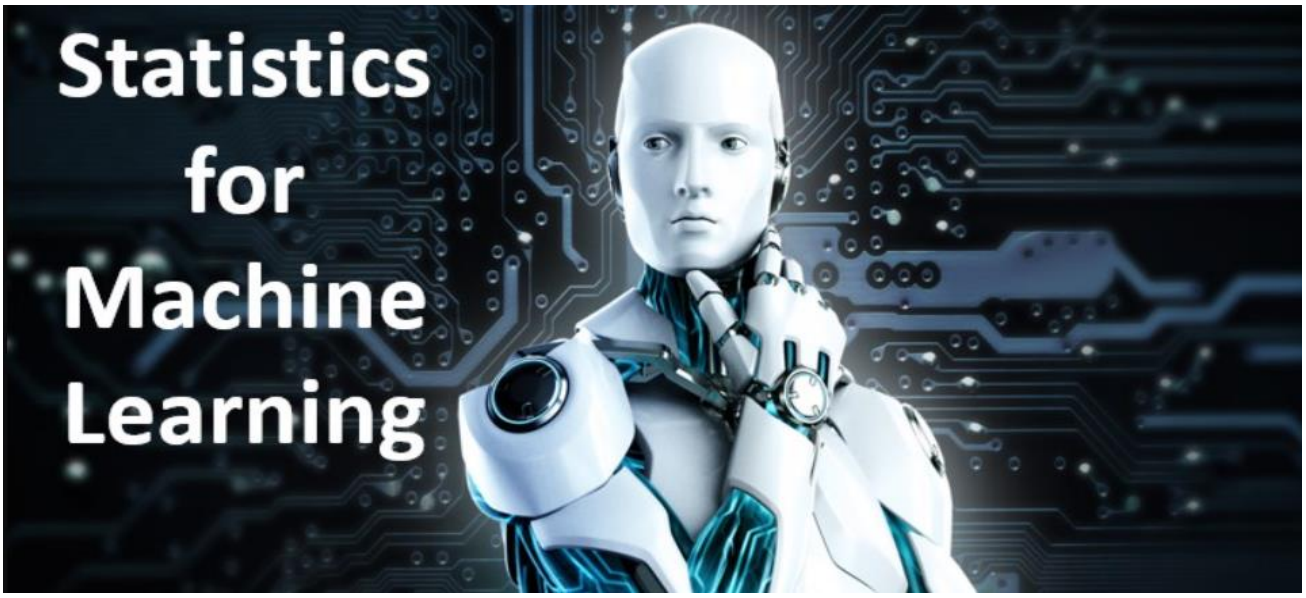
Statistics for Machine Learning Part 01



1. What are the most important topics in statistics for machine learning?

Breakdown of the most important topics in statistics, including explanations and additional terms:

Foundations:

- **Probability:** The likelihood of an event occurring. It helps quantify uncertainty and make predictions.
- **Random Variables:** Variables whose values depend on chance or randomness (e.g., coin toss outcome).
- **Probability Distributions:** Mathematical descriptions of the probability of different outcomes for a random variable (e.g., normal distribution for height).
- **Population vs. Sample:** The entire group you're interested in (population) vs. a representative subset used for analysis (sample).

Descriptive Statistics:

- **Measures of Central Tendency:** Summarize the "centre" of the data (e.g., mean, median, and mode).
- **Measures of Dispersion:** Describe how spread out the data is (e.g., variance, standard deviation, range).

Relationships between Variables:

- **Covariance:** Measures the direction and strength of the linear relationship between two variables.
- **Correlation:** Similar to covariance, but expresses the relationship as a single value between -1 (perfect negative) and 1 (perfect positive). Doesn't imply causation.

Statistical Inference:

- **Hypothesis Testing:** Formulating a null hypothesis (no effect) and alternative hypothesis (effect exists), then using data to assess evidence against the null hypothesis.
- **Confidence Intervals:** A range of values likely to contain the true population parameter, expressing the uncertainty around an estimate.
- **Central Limit Theorem:** Under certain conditions, the sampling distribution of averages (means) from a population approaches a normal distribution, regardless of the original population's shape (important for using samples to infer about populations).

Additional Important Topics:

- **Regression Analysis:** Modelling the relationship between a dependent variable (predicted) and one or more independent variables (predictors).
- **Sampling Techniques:** Methods for selecting representative samples (e.g., random sampling, stratified sampling).
- **Experimental Design:** Planning and conducting experiments to control variables and draw valid conclusions.
- **Power Analysis:** Determining the probability of detecting a true effect, given a specific sample size and design.
- **Bayesian Statistics:** An alternative approach to statistics that incorporates prior knowledge or beliefs into analysis.
- **Non-parametric Statistics:** Techniques used when data doesn't meet assumptions of traditional methods (e.g., median test, chi-square test).

This list is not exhaustive, but it covers the core concepts that form the foundation of statistical analysis. The specific topics you delve deeper into will depend on your field of study and the types of data you work with.

2. How Statistics Paves the Way for Machine Learning?

Imagine a treasure trove of information, but without a map to navigate it. That's where statistics comes in! It acts as the compass and key, helping us unlock the secrets hidden within data. But how does it connect to the marvels of machine learning (ML)? Let's delve into this fascinating partnership.

Statistics: The Bedrock of Data Analysis

Statistics provides a powerful toolkit for analysing data. It allows us to:

- **Uncover Patterns:** By summarizing and visualizing data through measures like mean, median, and standard deviation, we can identify trends and relationships that might not be readily apparent. This is crucial for training ML algorithms, as they learn from patterns.
- **Quantify Uncertainty:** Data isn't always perfect, and statistics helps us understand the inherent variability. Techniques like variance and standard deviation tell us how "spread out" the data is, giving valuable insights for building robust ML models that can handle real-world complexities.
- **Draw Meaningful Conclusions:** Through techniques like hypothesis testing, we can assess the strength of evidence within data. This helps us interpret the results of ML models and determine if they've truly learned valuable patterns, or if it's just a lucky coincidence.

Statistics as Fuel for Machine Learning

Machine learning algorithms are data-hungry beasts. They require vast amounts of high-quality data to learn and improve. Statistics plays a vital role in this process:

- **Data Cleaning and Preparation:** Before feeding data to an ML model, it needs to be cleaned and prepped. Techniques like outlier detection and feature scaling (using statistical measures) ensure the data is in a format the algorithm can understand and utilize effectively.
- **Model Selection and Evaluation:** Statistics helps us choose the right ML model for the job. By understanding the types of data and desired outcomes, we can select models that are statistically sound and well-suited to the task. Evaluating the performance of an ML model also relies on statistical methods, allowing us to assess its accuracy, precision, and potential for bias.

The Flip Side: Limitations of Statistics

While statistics is a powerful tool, it's not without limitations:

- **Data Dependence:** Statistical methods rely heavily on the quality and representativeness of the data they analyse. Garbage in, garbage out! If the data is biased or flawed, the statistical insights and resulting ML models can be misleading.
- **Assumptions and Oversimplification:** Statistical methods often make assumptions about the underlying data distribution. If these assumptions aren't met, the results can be unreliable. ML models, trained on these unreliable results, might perform poorly or produce inaccurate predictions.

The Takeaway: A Beautiful Partnership

Statistics and machine learning are like peanut butter and jelly. Together, they create something far greater than the sum of their parts. Statistics provides the foundation for understanding data, while machine learning leverages that understanding to build powerful models. By acknowledging the limitations of both, we can create a robust and insightful journey through the world of data.

3. The Power of Statistics: Shaping Decisions and Driving Progress

Statistics isn't just about numbers; it's about unlocking the stories data tells. From planning business strategies to guiding government policies, statistics plays a crucial role in various fields. Let's explore its importance and vast scope with some real-world applications:

1. Planning: A Data-Driven Compass

Imagine launching a new product without understanding market trends or customer preferences. Risky, right? Statistics equips us with the knowledge to make informed decisions through planning. Here's how:

- **Business:** Companies use statistical analysis of sales data to forecast demand, optimize inventory management, and target marketing campaigns effectively.
    - **Example:** A clothing retailer analyses past sales data to predict popular styles and sizes for the upcoming season, minimizing the risk of overstocking unpopular items.
- **Economics:** Governments rely on statistical data on factors like inflation, unemployment, and GDP growth to formulate economic policies that stimulate growth and stability.

    - **Example:** Analysing unemployment data helps policymakers identify sectors with high job losses, allowing them to create targeted economic stimulus programs.
- **Government:** Statistics are essential for planning public infrastructure projects. Analysing population data helps determine the need for schools and hospitals in specific areas.
    - **Example:** Traffic data analysis can be used to identify congested areas, informing decisions about road expansions or public transportation improvements.

2. Statistics: The Language of Mathematics

Statistics and mathematics go hand-in-hand. Mathematical concepts like probability theory form the foundation of statistical analysis. This powerful union has led to the development of:

- **Mathematical Statistics:** This branch delves deeper into the mathematical underpinnings of statistical methods, ensuring their accuracy and reliability.

Real-World Example: In clinical trials for new drugs, statistical methods are used to determine the effectiveness and safety of the drug by analysing data from test subjects.

3. Economics: The Language of Numbers

Statistics and economics are two sides of the same coin. Here's how statistics empowers economic decision-making:

- **Understanding Consumer Behaviour:** Businesses and governments use statistical analysis of consumer spending patterns (income distribution) to develop targeted economic policies.
    - **Example:** By analysing income data, policymakers can identify demographics that require social safety nets like food stamps or tax breaks, promoting economic equality.
- **Measuring Economic Health:** Statistical methods track key economic indicators like GDP growth, inflation, and unemployment rates. These metrics guide decisions on interest rates, government spending, and overall economic stability.
    - **Example:** Central banks analyse inflation data to determine if they need to raise interest rates to curb inflation, which can harm the economy.

4. Social Sciences: Unveiling the Human Experience

Statistics plays a crucial role in understanding social phenomena:

- **Identifying Trends:** Regression analysis helps isolate the impact of various factors on social issues like crime rates or educational attainment.
    - **Example:** Sociologists might use regression analysis to understand how factors like poverty and family structure influence teen pregnancy rates.
- **Population Studies:** Statistics are essential in demographics, the study of population trends. This data informs policies on healthcare, education, and social security.
    - **Example:** Governments use population growth data to predict future needs for schools and hospitals, ensuring adequate resources for citizens.

5. Trade: Navigating Uncertainty with Data

In the world of business, statistics are a compass for navigating uncertainty:

- **Market Forecasting:** Businesses use statistical analysis of historical sales data and market trends to forecast future demand for products and services.
    - **Example:** A clothing retailer might analyze past sales data to predict the demand for winter jackets in different regions, optimizing inventory management.
- **Price Optimization:** Statistical methods help businesses set optimal prices for products by considering factors like cost of production, competitor pricing, and customer demand.
    - **Example:** E-commerce companies use statistical techniques like A/B testing to determine the most effective product prices that maximize sales.

6. Research: The Foundation of Discovery

Statistics are the bedrock of rigorous research across various disciplines:

- **Data Analysis and Interpretation:** Researchers use statistical methods like hypothesis testing to analyze data collected through experiments or surveys, drawing conclusions about their research questions.
    - **Example:** A medical researcher might use statistical analysis to determine if a new drug treatment is effective in reducing a specific disease.

7. Healthcare: Statistics for a Healthier Future

Statistics are the backbone of evidence-based medicine, ensuring better patient care:

- **Clinical Trials:** Statistical analysis helps determine the effectiveness and safety of new drugs and treatments during clinical trials.
    - **Example:** Researchers use statistical methods to compare a new drug's effectiveness against a placebo, ensuring patients receive safe and beneficial treatments.
- **Disease Surveillance:** Statistics are used to track and analyse disease outbreaks, allowing for early intervention and control of epidemics.
    - **Example:** Public health officials use statistical analysis of disease outbreaks to identify hotspots and implement targeted prevention measures like vaccination campaigns.
- **Medical Research:** Statistical methods analyse data from medical studies, helping researchers understand disease risk factors and develop effective treatment strategies.
    - **Example:** Researchers might use statistical analysis to identify genetic mutations that increase the risk of certain cancers, leading to the development of targeted therapies.

8. Energy Sector: Powering Efficiency with Data

Statistics are crucial for optimizing energy production, distribution, and consumption:

- **Demand Forecasting:** Statistical analysis of historical energy usage patterns helps predict future demand, ensuring sufficient energy production and avoiding shortages.
    - **Example:** Energy companies use statistical models to predict electricity demand during peak seasons like summer, allowing them to adjust power generation accordingly.
- **Resource Exploration and Management:** Statistical methods are used to analyse geological data to locate and assess potential oil, gas, and renewable energy resources.
    - **Example:** Geologists use statistical techniques to analyse seismic data to identify areas with promising oil and gas reserves.
- **Energy Efficiency Analysis:** Statistical methods help identify areas for energy conservation and optimize energy usage in buildings and industries.
    - **Example:** Building managers use statistical analysis of energy consumption data to identify areas where they can implement energy-saving measures like improving insulation or upgrading lighting systems.

4. Measures of Central Tendency: Mean, Median, and Mode

A measure of central tendency is a summary statistic that represents the center point or typical value of a dataset. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution. You can think of it as the tendency of data to cluster around a middle value. In statistics, the three most Common measures of central tendency are the mean, median, and mode. Each of these measures calculates the location of the central point using a different method. The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriateto be used.

MEAN (ARITHMETIC)

The mean (or average) is the most popular and well-known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data (see our Types of Variable guide for data types). The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values X1, X2, X3… Xn , the sample mean, usually denoted by (pronounced x bar), is:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

This formula is usually written in a slightly different manner using the Greek capital letter, , pronounced "sigma", which means "sum of...".

$$\bar{x} = \frac{\sum x}{n}$$

You may have noticed that the above formula refers to the sample mean. So, why have we called it a sample mean? This is because, in statistics, samples and populations have very different meanings and these differences are very important,

even if, in the case of the mean, they are calculated in the same way. To acknowledge that we are calculating the population mean and not the sample mean, we use the Greek lowercase letter “mu", denoted as μ:

$$\mu = \frac{\sum x}{n}$$

The mean is essentially a model of your data set. It is the value that is most common. You will notice, however, that the mean is not often one of the actual values that you have observed in your data set. However, one of its important properties is that it minimizes error in the prediction of any one value in your data set. That is, it is the value that produces the lowest amount of error from all other values in the data set.

An important property of the mean is that it includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

MEDIAN

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below:

65	55	89	56	35	14	56	55	87	45	92

We first need to rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	**56**	56	65	87	89	92

Our median mark is the middle mark — in this case, 56 (highlighted in bold). It is the middle mark because there are 5 scores before it and 5 scores after it. This works fine when you have an odd number of scores, but what happens when you have an even number of scores? What if you had only 10 scores? Well, you simply have to take the middle two scores and average the result. So, if we look at the example below:

14	35	45	55	**55**	**56**	56	65	87	89

Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.

Example of Median:

1, 3, 3, **6**, 7, 8, 9

Median = 6

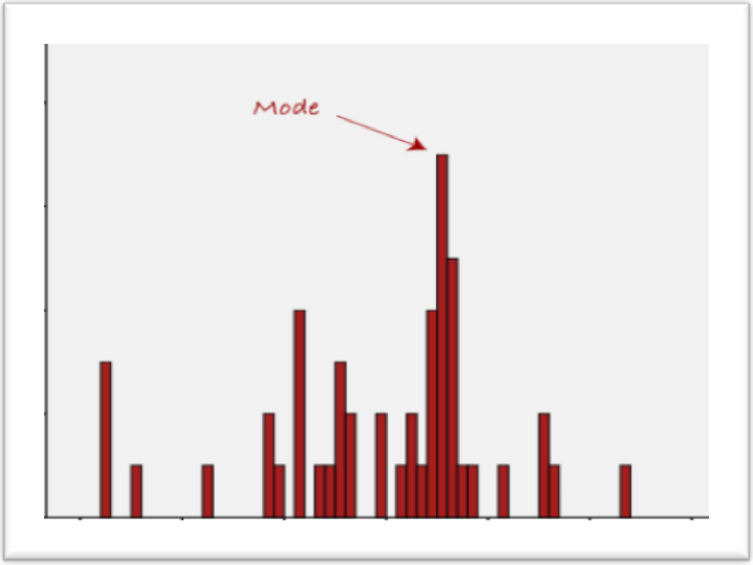
1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= 4.5

MODE

The mode is the most frequent score in our data set. On a histogram, it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option. An example of a mode is presented below:

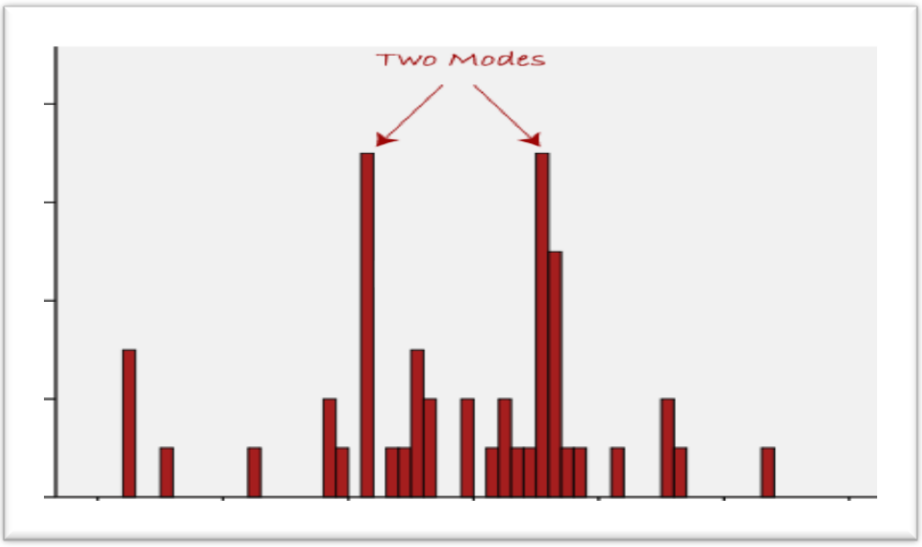


Why is Mode Rarely used with Continuous data?

The mode is particularly problematic with continuous data because it is more likely not to have any value that is more frequent than the other. For example, consider the data set consisting of the weights of 30 people. How likely is it that that two or more people with exactly the same weight (e.g., 55.4 kg) are present in the same sample? The answer would be that it is perhaps highly unlikely. Though many people might be close, it is impossible to find two people with exactly the same weight (to the nearest 0.1 kg), with such a small sample (30 people) and a large range of possible weights. This is why the mode is very rarely used with continuous data.

Other Limitations of Using Mode

One of the major limitations with the mode is that it is not unique. So it leaves with problems when having two or more values that share the highest frequency, such as below:



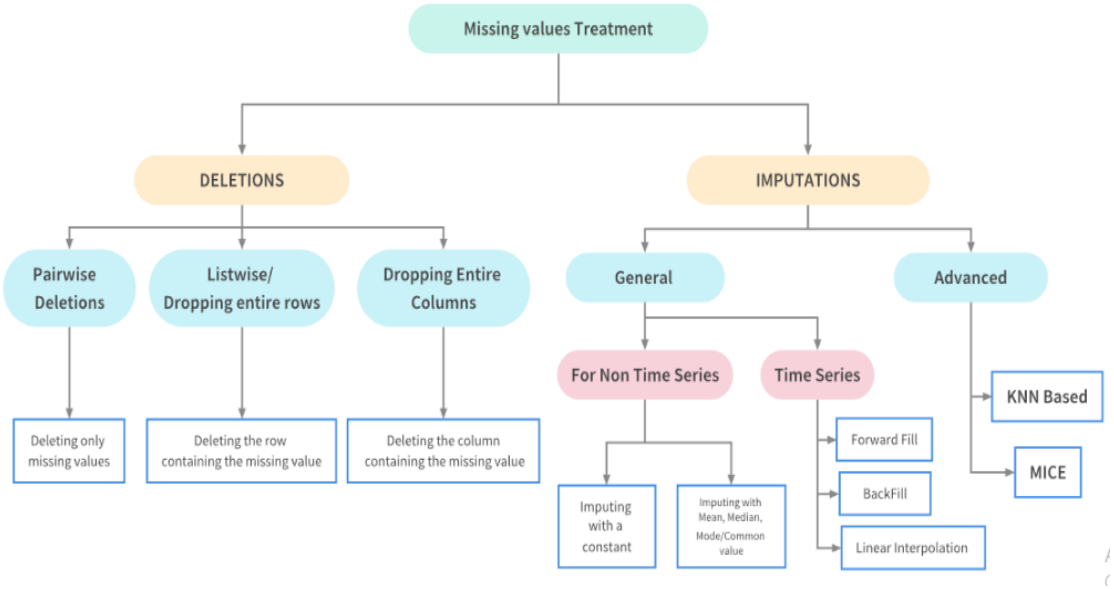
Summary of When to Use Mean, Median and Mode

The below table will help to choose the best measures of central tendency with respect to different types of variables.

Type of Variable	The Best Measure of Central Tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

5. How would you approach a dataset that's missing more than 30% of its values?

There are several approaches you can take when dealing with a dataset missing more than 30% of its values, depending on the specific situation and your goals for the analysis. Here are some options to consider:



1. Removal:

- **Listwise Deletion:** If the missing data is randomly distributed and a significant portion (more than 30%) is missing, simply removing rows or columns with missing values might be the quickest solution. However, this approach can lead to a loss of information and potentially bias your results, especially if the missing data is not random.

2. Imputation:

- This involves estimating the missing values based on the available data. There are various imputation techniques, each with its own advantages and limitations:
    - **Mean/Median/Mode Imputation:** Replace missing values with the average (mean), middle value (median), or most frequent value (mode) of the existing data in that column. This is a simple method but might not be suitable for skewed data or if the missing values are not randomly distributed.
    - **Interpolation:** Estimate missing values based on surrounding data points. This can be linear interpolation (connecting adjacent values with a straight line) or more complex methods depending on the data.
    - **Model-based Imputation:** Use statistical models like regression analysis to predict missing values based on the relationships between variables. This can be effective but requires careful model selection and validation.

3. Dimensionality Reduction:

- If you have a large number of features (columns) and many missing values, consider reducing the dimensionality of your data. Techniques like Principal Component Analysis (PCA) can help identify underlying patterns and create new features with less missing data.

4. Alternative Analysis Techniques:

- Depending on your research question, there might be alternative statistical methods that can handle missing data more effectively. For example, robust statistics are less sensitive to outliers and missing values.

Here are some additional factors to consider when choosing an approach:

- **The nature of the missing data:** Is it missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR)? This can influence the validity of different imputation techniques.
- **The impact on your analysis:** How much does the missing data affect the accuracy and reliability of your results?
- **The importance of preserving all data points:** If retaining all data points is critical, imputation might be a better option than removal.

It's important to be transparent about the chosen approach and the potential limitations it introduces when presenting your analysis.

Ultimately, the best approach depends on the specific dataset and your research question.

6. Give an example where the median is a better measure than themean.

Outliers and skewed data often push the mean to the extreme, making the median a better indicator of where most values cluster in the dataset

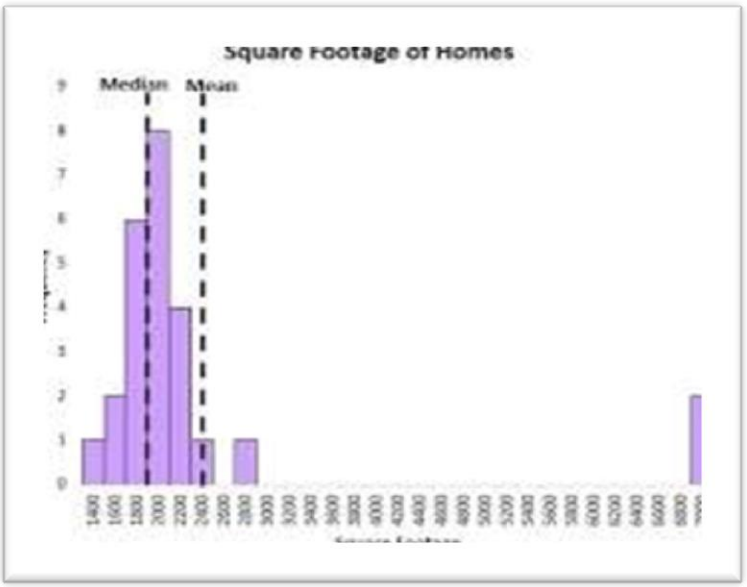
Example: Income Distribution

Suppose we're analysing the income of a group of people. Here's why the median might be a better measure:

1. **Skewed Data**: Income distributions are often skewed, with a few extremely high earners. If we calculate the mean income, these outliers significantly impact the result.

The mean gets pulled toward the higher values, making it less representative of the typical person's income.

2. **Robustness:** The median is a robust statistic. It's not affected by extreme values (outliers) as much as the mean. Even if a few people have exceptionally high or low incomes, the median remains relatively stable.
3. **Interpretability**: The median represents the middle value when data is sorted. For income, it's the income level at which half the population earns more and half earns less. This intuitively captures the "typical" income.



Example Calculation:

Suppose we have the following income data (in thousands of dollars):

Incomes: 20,25,30,40,50,1000

- **Mean Income**:

  Mean=620+25+30+40+50+1000= 191.67 thousand dollars

- **Median Income**: Sort the incomes:

  20,25,30,40,50,1000

  Median = 35 (the middle value)

In this case, the median ($35,000) better represents the typical income compared to the mean ($191,670). It's less affected by the outlier ($1,000,000).

Remember, context matters, and choosing the right measure depends on the specific characteristics of the data!

7. What is the difference between Descriptive and Inferential Statistics?

Aspect	Descriptive Statistics	Inferential Statistics
Purpose	Describes a dataset using summary statistics, graphs, and tables.	Makes predictions or generalizations about a larger dataset based on a sample.
Focus	Visible characteristics of a dataset (population or sample).	Likelihood of future event occurrence.
Examples	Mean, median, range, standard deviation, histograms, frequency tables.	Hypothesis testing, confidence intervals, regression, ANOVA.
Representation	Summary statistics, graphs, and tables.	Displayed in the form of probability.
Application Scenario	Understanding data quickly, visualizing distributions, identifying central tendencies and variability.	Drawing conclusions about populations based on samples, assessing treatment effects, predicting outcomes.

Descriptive and Inferential Statistics are two fundamental tools used to analyse data, but they serve different purposes. Here's a breakdown of their key differences:

Focus:

- **Descriptive Statistics:** Focuses on describing and summarizing the characteristics of a **single dataset**. It provides a snapshot of the data without making generalizations about a larger population.
- **Inferential Statistics:** Uses sample data to draw inferences about a **larger population**. It allows us to estimate population parameters and test hypotheses about relationships between variables.

Data Used:

- **Descriptive Statistics:** Can be applied to the entire dataset (**population**) or a **sample**. However, it's most informative when used with the entire dataset.
- **Inferential Statistics:** Relies on **sample data** drawn from the population of interest.

Goal:

- **Descriptive Statistics:** Aims to organize, summarize, and present data in a way that's easy to understand. It helps identify patterns, trends, and central tendencies within the data itself.
- **Inferential Statistics:** Goes beyond the data at hand. It allows us to make predictions or generalizations about a larger population based on the information gathered from the sample.

Techniques:

- **Descriptive Statistics:** Uses various methods to summarize data, including:
    - Measures of central tendency (mean, median, mode)
    - Measures of dispersion (variance, standard deviation)
    - Graphical representations (histograms, boxplots)
- **Inferential Statistics:** Employs more complex techniques to draw inferences, such as:
    - Hypothesis testing
    - Confidence intervals
    - Regression analysis

Example:

- **Descriptive Statistics:** Calculating the average exam score and the distribution of grades in a class provides a descriptive picture of student performance.
- **Inferential Statistics:** Using a sample of voters to estimate voter preferences in an entire city allows us to infer the likely outcome of the election for the entire population.

Certainty:

- **Descriptive Statistics:** Offers a high degree of certainty about the data being analysed since it's based on the entire dataset (ideally) or a representative sample.
- **Inferential Statistics:** Results have a margin of error due to using samples. There's always some level of uncertainty when making inferences about a larger population.

Application:

- **Descriptive Statistics:** Provides a foundational understanding of data, helping us identify patterns and trends within a dataset. This is crucial for data exploration and visualization.
- **Inferential Statistics:** Allows us to make informed decisions based on broader population trends beyond the data immediately available. It's valuable for hypothesis testing, prediction, and understanding relationships between variables.

In essence, Descriptive Statistics paints a picture of the data itself, while Inferential Statistics uses that picture to make educated guesses about the bigger picture - the entire population.

8. How Can We Slice Up Our Data? Exploring Quartiles, Deciles, and Percentiles

Partition values or fractiles such a quartile, a decile, etc. are the different sides of the same story. In other words, these are values that divide the same set of observations in different ways. So, we can fragment these observations into several equal parts.

QUARTILE

A quartile is a statistical value that divides a sorted dataset into four equal parts. It helps us understand how the data is distributed and where most of the values fall.

Concept:

- Imagine you have a list of exam scores for a class, arranged from lowest to highest.
- Dividing this list into four equal parts gives you three quartile values: Q1, Q2, and Q3.

Types of Quartiles:

- **First Quartile (Q1):** Also called the lower quartile, it represents the value at which 25% of the data falls below it and 75% falls above it.
- **Second Quartile (Q2):** This is the median of the dataset. It represents the middle value when the data is sorted, with 50% of the data falling below it and 50% above it.
- **Third Quartile (Q3):** Also called the upper quartile, it represents the value at which 75% of the data falls below it and 25% falls above it.

Applications:

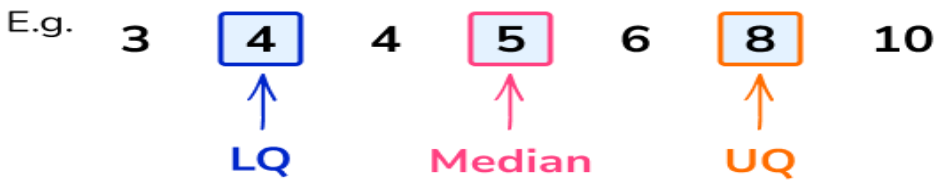
- Understanding the spread of data: Knowing the quartiles gives you an idea of how spread out the data is. A small difference between Q1 and Q3 indicates that most of the data is clustered around the median (Q2). A large difference suggests a wider spread with more values towards the extremes.
- Identifying outliers: Values significantly lower than Q1 or higher than Q3 can be potential outliers that deserve further investigation.
- Comparing datasets: Quartiles allow you to compare the distribution of data across different groups or populations.

Quartiles divide an ordered data set into four equal parts (quarters). We use subscript notation to label the quartiles: Q1, Q2 and Q3.

The first quartile, Q1, is $\frac{1}{4}$ (or 25%) of the way through the data - the **lower quartile**.

The second quartile, Q2 is $\frac{1}{2}$ (or 50%) of the way through the data - the **median**.

The third quartile, Q3 is $\frac{3}{4}$ (or 75%) of the way through the data - the **upper quartile.**



DECILES

Deciles are statistical values that divide a sorted dataset into ten equal parts. They provide more granularity compared to quartiles (which divide data into fourths).

Number of Deciles:

There are actually **nine deciles** (D1 to D9), not eight. Each decile marks a point where a specific percentage of the data falls below it.

D1 Explained:

D1 refers to the **first decile**. It's not the "typical peak value." Here's a breakdown:

- **D1:** Represents the value at which 10% of the data falls below it and 90% falls above it.

Example: Imagine a dataset containing exam scores for 100 students, arranged from lowest to highest.

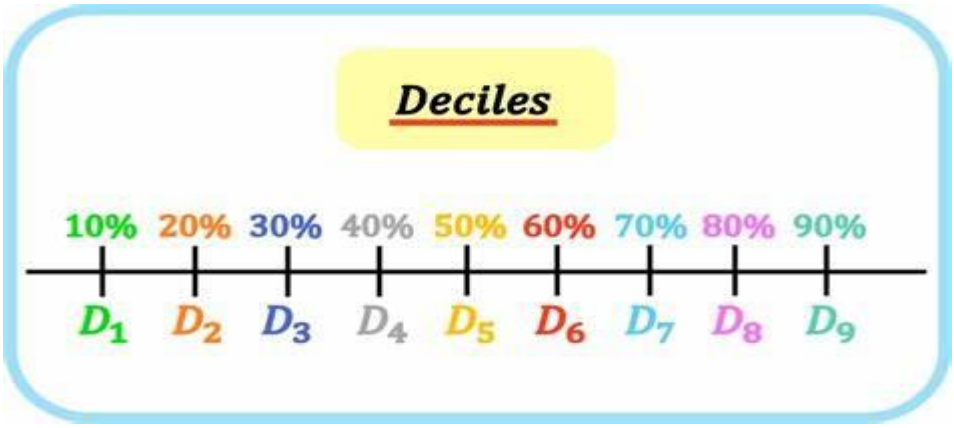
- D1 would be the score at which 10 students (10%) scored lower than or equal to that value. The remaining 90 students scored higher than D1.

Understanding Deciles with the Example:

- D2: The score where 20% of students scored lower and 80% higher.
- D3: The score where 30% scored lower and 70% higher, and so on.
- D9: The score where 90% of students scored lower and 10% higher (almost reaching the highest score).

Applications of Deciles:

- **Detailed Distribution:** Deciles provide a more precise picture of how data is spread out compared to quartiles.
- **Identifying Specific Segments:** They can be used to identify specific segments within a dataset. For example, D5 (median) divides the data into half, while D7 might represent the score above which only the top 30% of students scored.



PERCENTILES

Percentiles, in statistics, are a generalization of quartiles and deciles. They represent specific values in a sorted data set that divide the data into equal parts. Here's a breakdown of percentiles:

Concept:

- Imagine a dataset like exam scores, arranged from lowest to highest.
- A percentile (denoted as Pn) divides the data into 100 equal parts. There are 99 percentiles (P1 to P99) because the 100th percentile would simply be the highest value itself.

Types of Percentiles:

- **P1:** Represents the value where 1% of the data falls below it and 99% falls above it.
- **P25:** This is the first quartile (Q1). It represents the value at which 25% of the data falls below it and 75% falls above it.
- **P50:** This is the median (Q2) of the dataset. It represents the middle value with 50% of the data falling below it and 50% above it.
- **P75:** This is the third quartile (Q3). It represents the value at which 75% of the data falls below it and 25% falls above it.
- **P99:** Represents the value where 99% of the data falls below it and 1% falls above it.

Applications:

- **Understanding Data Distribution:** Percentiles provide a detailed picture of how data is spread out. Knowing percentiles like P10 and P90 tells you where the lowest 10% and highest 10% of the data points lie, respectively.
- **Identifying Outliers:** Values significantly lower than a low percentile (like P1) or higher than a high percentile (like P99) can be potential outliers.
- **Comparing Datasets:** Percentiles allow you to compare the distribution of data across different groups or populations. For example, comparing the P50 (median) of exam scores for two classes can reveal which class performed better overall.
- **Standardized Tests:** Many standardized tests report scores in percentiles. A score at the 80th percentile indicates that the test-taker performed better than 80% of the people who took the same test.

Choosing the Right Percentiles:

- Quartiles (P25, P50, P75) are commonly used for a basic understanding of data spread.
- Deciles (P10, P20, ..., P90) provide more detailed information.

- You can choose any specific percentile (like P90 or P95) to pinpoint a specific point in the data distribution.

In essence, percentiles offer a flexible way to analyse data distribution, allowing you to zoom in on specific segments or get a broad picture of how the data is spread out.

Calculation Steps:

- **Order the Data**: Arrange the data in ascending order.
- **Assign Ranks**: Assign a rank to each data value (1 for the smallest, 2 for the next, and so on).
- **Calculate Ordinal Rank**: To find a specific percentile, compute the ordinal rank using this formula:

  $$\text{Ordinal Rank} = (100\text{percentile}) \times \text{total number of data points}$$

- **Identify the Value**: The value with the next rank after the ordinal rank is the desired percentile value.

Example 1:

- Data: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
- Percentiles:
  - 30th percentile:
    $\text{Ordinal Rank} = (100 30) \times 10 = 3$
    (Next rank: 4, value: 40)
  - 40th percentile:
    $\text{Ordinal Rank} = (100 40) \times 10 = 4$
    (Next rank: 5, value: 50)
  - 50th percentile:
    $\text{Ordinal Rank} = (100 50) \times 10 = 5$
    (Next rank: 6, value: 60)

  - 100th percentile: Assume 100 is the maximum value.

Example 2 (Ages in years):

- Data: 25, 25, 26, 36, 39, 40, 40, 44, 44, 44, 45, 47, 48, 51, 52, 52, 52, 53, 67, 77

- Percentiles:
  - 10th percentile: 25
  - 30th percentile: 36
  - 60th percentile: 52
  - 80th percentile: 52

9. How Spread Out Is Our Data? Exploring Range and Interquartile Range (IQR)?

We've learned about measures of central tendency, like mean or median, which tell us the typical value in a dataset. But data isn't always perfectly clustered around that central point. Some datasets are spread out widely, while others are tightly packed together. This spread or dispersion is just as important to understand as the average value. In statistics, we have different terms for this spread: variability, dispersion, or simply how "spread out" the data is. Similar to having multiple ways to measure the centre, there are also several ways to quantify how spread out the data points is in a distribution.

Key Points:

- Measures of central tendency (mean, median, and mode) describe the "typical" value in a dataset.
- Measures of variation tell us how spreads out the data points are around the central value.
- A low variation indicates the data points are close to the centre.
- High variation signifies the data points are scattered further away from the centre.
- Variability, dispersion, and spread all refer to the same concept - how wide the distribution of data is.

RANGE

Let's start with the range because it is the most straightforward measure of variability to calculate and the simplest to understand. The range of a dataset is the difference between the largest and smallest values in that dataset. For example, in the two datasets below, dataset 1 has a range of 20 — 38 = 18 while dataset 2 has a range of 11 — 52 = 41. Dataset 2 has a wider range and, hence, more variability than dataset 1.

Dataset 1	Dataset 2
20	11
21	16
22	19
25	23
26	25
29	32
33	39
34	46
38	52

While the range is easy to understand, it is based on only the two most extreme values in the dataset, which makes it very susceptible to outliers. If one of those numbers is unusually high or low, it affects the entire range even if it isatypical.

Additionally, the size of the dataset affects the range. In general, you are less likely to observe extreme values. However, as you increase the sample size, you have more opportunities to obtain these extreme values.

Consequently, when you draw random samples from the same population, the range tends to increase as the samplesize increases. Consequently, use the range to compare variability only when the sample sizes are similar.

THE INTERQUARTILE RANGE (IQR)

The interquartile range is the middle half of the data. To visualize it, think about the median value that splits the dataset in half. Similarly, you can divide the data into quarters. Statisticians refer to these quarters as quartiles and denote them from low to high as Q1, Q2, Q3, and Q4. The lowest quartile (Q1) contains the quarter of the dataset with the smallest values. The upper quartile (Q4) contains the quarter of the dataset with the highest values. The interquartile range is the middle half of the data that is in between the upper and lower quartiles. In other words, the interquartile range includes the 50% of data points that fall in Q2.

Example:

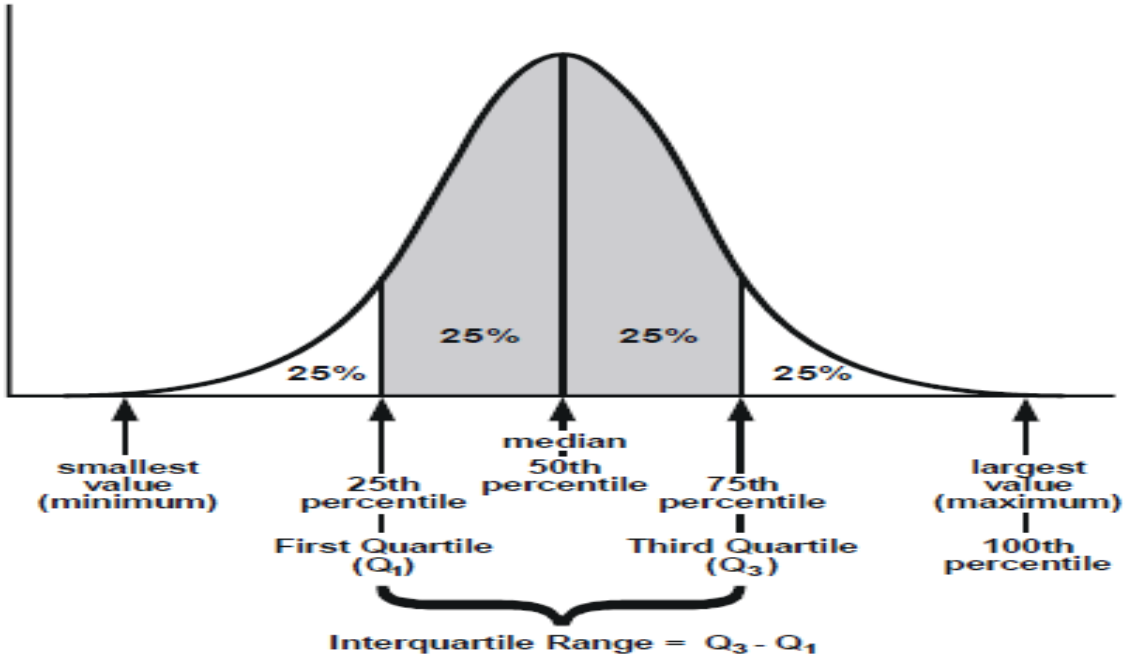
Imagine you have a bunch of leaves and want to know how spreads out their sizes are. The average (mean) size tells you one thing, but it doesn't reveal the whole picture. The interquartile range (IQR) helps us understand the "middle majority" of leaf sizes.

Here's how it works:

1. **Splitting the Data:** We can divide the leaves (or any data set) into four equal quarters, ordered from smallest to largest. Statisticians call these quarters "quartiles" and label them Q1 (lowest), Q2 (median), Q3 (highest), and (confusingly) not Q4!
2. **Focusing on the Middle:** The IQR specifically zooms in on the **middle two quartiles** (Q2 and Q3). **Q2**, the median, represents the exact middle value.
3. **IQR: The Gap in the Middle:** The IQR is the difference between the value in the upper quartile (Q3) and the value in the lower quartile (Q1). In simpler

terms, it tells you the range of values that encompasses the middle 50% of the data, excluding the most extreme values at either end.

Key takeaway: The IQR provides a clearer picture of how spread out the majority of the data points are, focusing on the central area and potentially revealing outliers on the fringes.



The **interquartile range** is the difference between the upper quartile and the lower quartile in a data set.

Statistic	Value
Lowest Value	3
Lower Quartile Q_1	4
Median M	5
Upper Quartile Q_3	8
Highest Value	10

$$IQR = UQ - LQ$$

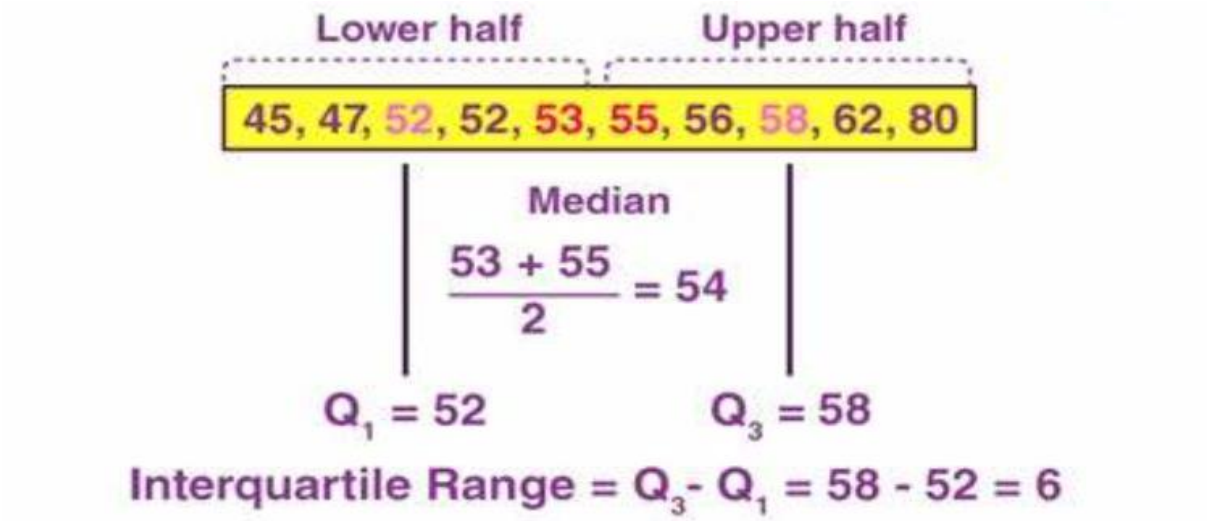
3 **4** 4 **5** 6 **8** 10

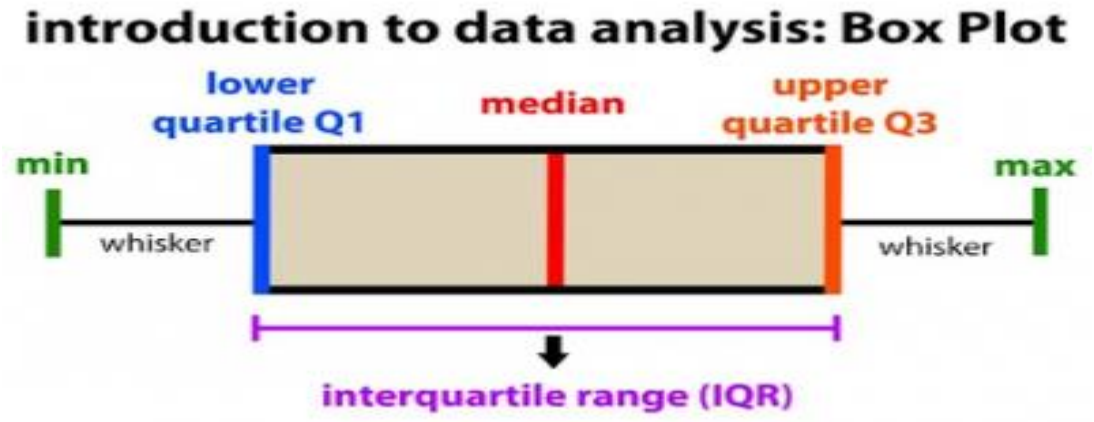
LQ M UQ

For this data set, $IQR = UQ - LQ = 8 - 4 = 4$

The five-number summary for this data is:

IQR Calculations





IQR: Taming Outliers and Skewed Data

We've learned that the median is a good central tendency measure because it's not easily swayed by extreme values (outliers). The interquartile range (IQR) shares this strength, making it a robust measure of variability.

Here's why IQR is less affected by outliers:

- **Focuses on the Middle:** Unlike measures like range (highest value minus lowest value), IQR concentrates on the **middle 50%** of the data, excluding the most extreme points on either side. Outliers have less influence on this central area.
- **IQR for Skewed Distributions:** Some datasets are skewed, meaning they have more data points clustered on one side. The standard deviation, which relies heavily on the mean, can be misleading in such cases. IQR, however, remains a reliable measure because it doesn't depend on every single data point and focuses on the central portion of the distribution.

In essence, IQR is a powerful tool for understanding data spread, especially when outliers or skewed distributions might distort other measures of variability.

10. Mean deviation and Standard deviation

Both mean deviation (MD) and standard deviation (SD) are measures of variability in statistics, but they differ in their approach and how they handle data. Here's a breakdown to understand their key differences:

Mean Deviation (MD):

- **Concept:** MD calculates the average distance of each data point from the mean (average) of the dataset. It reflects how spread out the data is around the central value.
- **Calculation:**
    1. Calculate the mean of the data.
    2. Find the absolute value of the difference between each data point and the mean.
    3. Take the average of those absolute differences.

Formula Mean Deviation (MD):

$$MD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \mu|$$

where:

- (n) is the number of data points.
- (x_i) represents each data value.
- (μ) is the mean of the data set.

Standard Deviation (SD):

- **Concept:** SD measures the average squared distance of each data point from the mean. It considers not just the direction (positive or negative difference) but also the magnitude of the deviation from the mean. Squaring the differences emphasizes larger deviations more heavily.
- **Calculation:**
    1. Calculate the mean of the data.
    2. Find the squared difference between each data point and the mean.
    3. Take the average of those squared differences.
    4. Finally, take the square root of the result obtained in step 3 (to bring the units back to the original scale of the data).

Formula Standard Deviation (SD):

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

where:

- (n) is the number of data points.
- (x_i) represents each data value.
- (μ) is the mean of the data set

Choosing Between MD and SD:

- **MD is less sensitive to outliers:** Since it uses absolute values, outliers have a smaller impact on MD compared to SD.
- **SD is more widely used and mathematically convenient:** It has many useful properties that make it easier to work with in statistical analysis.
- **MD is easier to interpret:** The units of MD are the same as the units of your data, making it easier to understand the spread in the original scale.
- **SD reflects the spread of squared distances:** While providing more weight to larger deviations, SD doesn't directly tell you how far individual points are from the mean (unlike MD).

In general:

- **Use SD when normality is assumed and outliers are not a major concern.**
- **Use MD when dealing with skewed data or when outliers might significantly affect the results.**