UNIT-1

INTRODUCTION TO STATISTICS

Data collection methods, Descriptive Statistics Mean, Median, Mode, Inferential Statistics, Random Variables, Probability Distributions, Normal Distribution, Sampling and Sampling Distribution.

Data :

- Data are “facts or information.
- Data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things.
-

Difference between the data and information:

Data: Data refers to raw, unprocessed facts and figures. It lacks context and meaning on its own. Data can be in the form of numbers, text, images, or any other format.

Example:

Name, Reg number, Phone number

Information: Information, on the other hand, is data that has been processed, organized, and interpreted within a specific context to make it meaningful and useful for decision-making or understanding. It provides insights, context, or significance to the raw data

Example:

Name and address is information.

Database:

- A database is an organized collection of structured information, or data, typically stored electronically in a computer system.
- The data can then be easily accessed, managed, modified, updated, controlled, and organized.

- Most databases use structured query language (SQL) for writing and querying data.
- Ex: Student Database, Employee Database

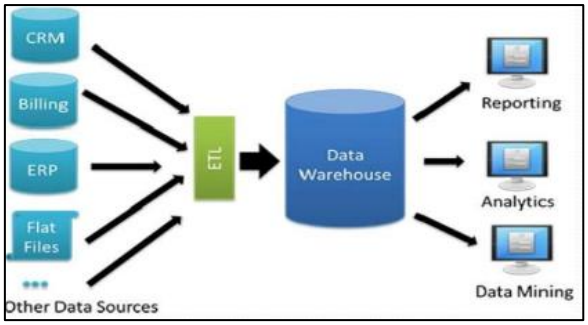
Data center:

- A Data center is a physical facility that organizations use to house their critical applications and data.
- A data center's design is based on a network of computing and storage resources that enable the delivery of shared applications and data
- The key components of a data center design include routers, switches, firewalls, storage systems, servers, and application-delivery controllers.

Data Warehouse:

- A data warehouse is a central repository of information that can be analyzed to make more informed decisions.
- Data flows into a data warehouse from transactional systems, relational databases, and other sources.
- Business analysts, data engineers, data scientists, and decision makers access the data through business intelligence (BI) tool

- Informed decision making
- Consolidated data from many sources
- Historical data analysis.
- Data quality, consistency, and accuracy. Separation of analytics
- processing from transactional databases, which improves performance of both systems.
- ETL- Extract, Transform and load



Data visualization:

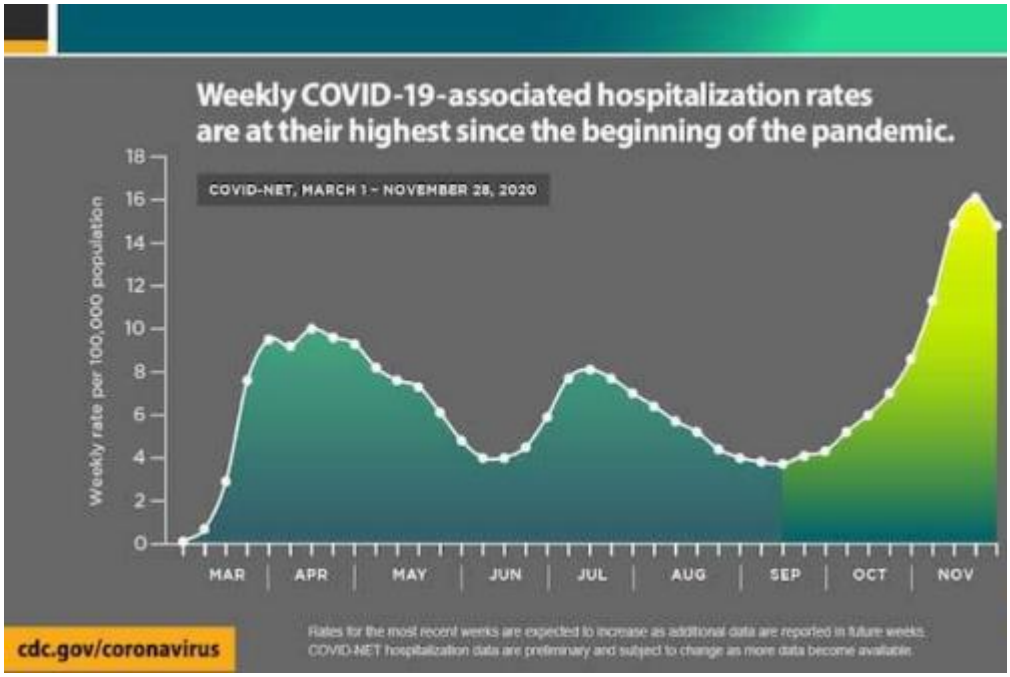
- In this technical world, a lot of data is being generated on a day by day.
- And sometimes to analyze this data for certain trends, patterns may become difficult if the data is in its raw format.
- Data visualization provides a good, organized pictorial representation of the data which makes it easier to understand, observe, and analyze.
- Communicating the information correctly too the required target.
- Data visualization is the graphical representation of information and data.

What Is Good Data Visualization?

- Accurate: The visualization should accurately represent the data and its trends.
- Clear: Your visualization should be easy to understand.
- Empowering: The reader should know what action to take after viewing your visualization.
- Succinct: Your message shouldn't take long to resonate.
- Data visualization is also an element of the broader data presentation architecture (DPA) discipline, which aims to identify, locate, manipulate, format and deliver data in the most efficient way possible.

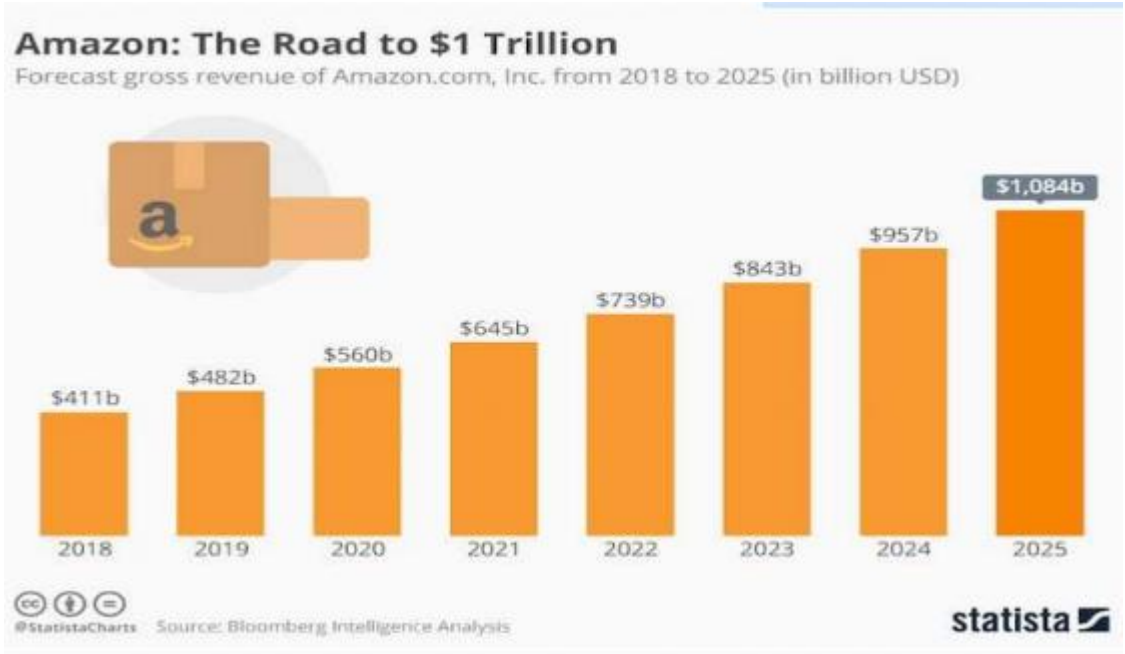
Real life examples:

COVID-19 Hospitalization Rates



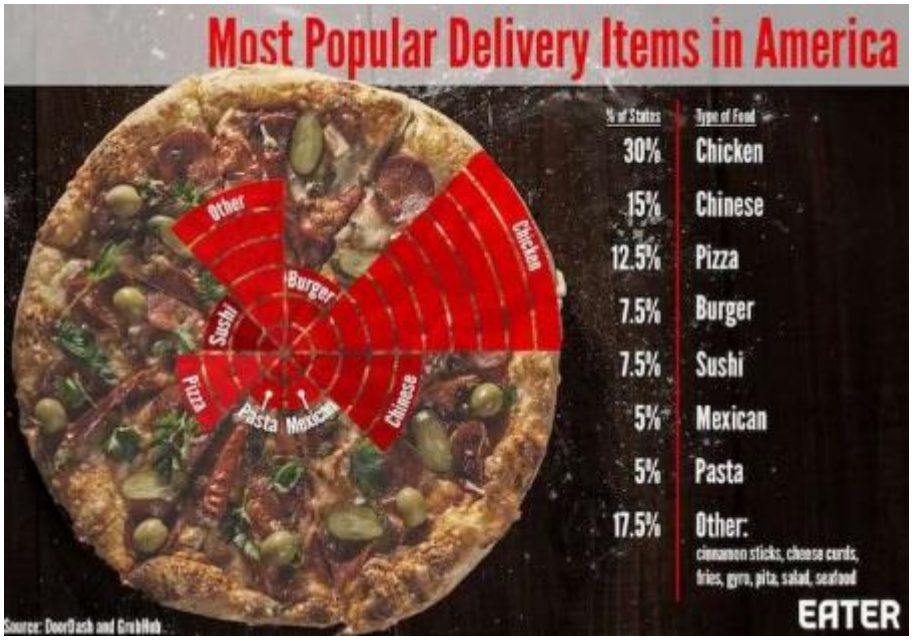
Revenue of Amazon.com:

Data visualizations aren't limited to historical data. This bar chart created by visualizes the forecasted gross revenue of Amazon.com from 2018 to 2025.



Most Popular Food Delivery Items

Dining brand has created this fun take on a “pie” chart, which shows the most common foods ordered for delivery in each of the United States.



Statistics:

- Statistics is a mathematical science including methods of collecting, organizing and analyzing data in such a way that meaningful conclusions can be drawn from them.
- Data can be defined as groups of information that represent the qualitative or quantitative attributes of a variable or set of variables.
- Statistics simply means numerical data, and is field of math that generally deals with collection of data, tabulation, and interpretation of numerical data.
- An example of data can be the ages of the students in a given class. When you collect those ages, that becomes your data.

why studying the field of statistics is crucial in modern society.

• The field of statistics is the science of learning from data. Statistical knowledge helps you use the proper methods to collect the data, employ the correct analyses, and effectively present the results.

• Statistics is a crucial process behind how we make discoveries in science, make decisions based on data, and make predictions. Statistics allows you to understand a subject much more deeply.

There are two main guidelines :

• First, statisticians are guides for learning from data and navigating common problems that can lead you to incorrect conclusions.

• Second, given the growing importance of decisions and opinions based on data, it's crucial that you can critically assess the quality of analysis that others present to you.

Data Collection:
Data collection in statistics refers to the process of gathering information or observations from various sources to be used for statistical analysis.

Type of data collection:

- Census data collection
- Sample data collection
- Experimental data collection
- Observational data collection

Census data collection:

Census data collection involves gathering comprehensive information from every individual or entity within a specific population or geographic area. It aims to provide a complete and accurate picture of the characteristics, demographics, and other relevant data of the entire population.

Example:

- India also conducts a decennial census to gather demographic information about its population.
- During the Indian census, which typically occurs every ten years, data is collected on various demographic factors such as population size, age distribution, sex ratio, literacy rate, religion, caste, languages spoken, employment status, housing conditions, and other socio-economic indicators.

Table 2.
States with the Fastest and Slowest Growth in Resident Population: 2010 to 2020

State	Population		Change	
	2010	2020	Number	Percent
Fastest Growing				
Utah	2,763,885	3,271,616	507,731	18.4
Idaho	1,567,582	1,839,106	271,524	17.3
Texas	25,145,561	29,145,505	3,999,944	15.9
North Dakota	672,591	779,094	106,503	15.8
Nevada	2,700,551	3,104,614	404,063	15.0
Slowest Growing				
Connecticut	3,574,097	3,605,944	31,847	0.9
Michigan	9,883,640	10,077,331	193,691	2.0
Ohio	11,536,504	11,799,448	262,944	2.3
Wyoming	563,626	576,851	13,225	2.3
Pennsylvania	12,702,379	13,002,700	300,321	2.4

Source: U.S. Census Bureau, 2020 Census and 2010 Census

Sample data collection:

Sample data collection, which is commonly just referred to as sampling, is a method which collects data from only a chosen portion of the population.

Real time Example:

Surveying to gauge public opinion on a political issue.

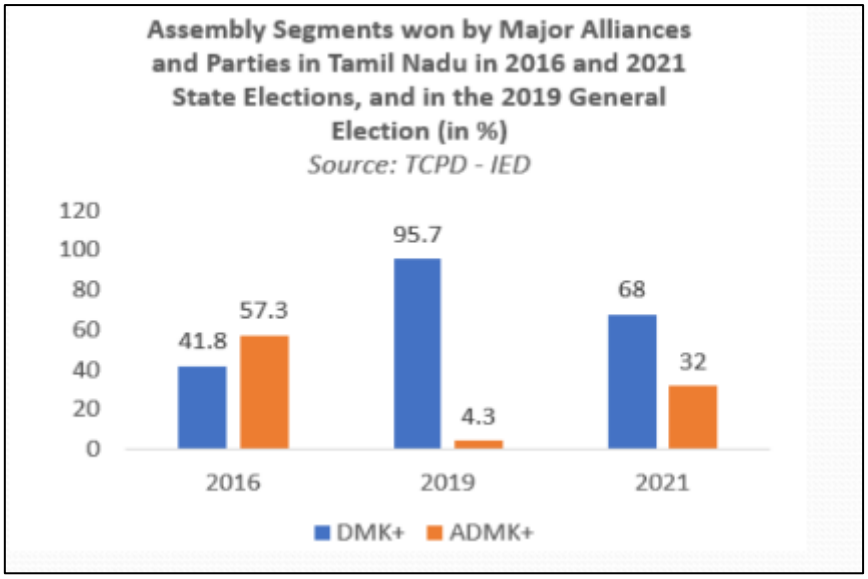
Polling organizations might select a representative sample of voters and ask them questions about their political preferences and opinions. The results from the sample are then used to make inferences about the broader population's opinions.

Market Research:

A company conducting a survey to understand consumer preferences for a new product. Market research firms often use sampling techniques to select a representative sample of consumers and ask them questions about their purchasing habits, brand preferences, product features, and pricing preferences. The data collected from the sample can then be analyzed to make informed decisions about product development, marketing strategies, and market positioning.

Health Surveys:

A public health organization conducting a survey to assess the prevalence of a particular disease or health risk factors within a population. Researchers might select a representative sample of individuals from different demographic groups and administer health questionnaires or conduct medical tests to collect data on health behaviors, chronic conditions, lifestyle factors, and access to healthcare services. The findings from the sample can provide valuable insights into the health needs of the population and inform public health interventions and policies.



Experimental Data collection

•Experimental data collection involves one performing an experiment and then collecting the data to be further analyzed. Experiments involve tests and the results of these Experiments that are conducted.

Example:

rolling a die onehundred times while recording the outcomes.

• Your data would be the results you get in each roll. The experiment could involve rolling the die in different ways and recording the results for each of those different ways

Testing a New Drug:

Testing a new drug to assess its effectiveness in treating a particular medical condition. Researchers would conduct a clinical trial where participants are randomly assigned to either receive the new drug

(treatment group) or a placebo (control group). By comparing the outcomes between the two groups, researchers can evaluate the efficacy of the drug. two more simple examples

Education Intervention Study:

Real-time Example: A study assessing the effectiveness of a new teaching method on student learning outcomes. Researchers may randomly assign classrooms or schools to either receive the new teaching method (experimental group) or continue with the existing method (control group). Pre-tests and post-tests could be administered to measure students' academic performance before and after the intervention. By comparing the improvement in test scores between the two groups, researchers can evaluate the impact of the new teaching method.

Observational data collection:

The observational data collection method involves not carrying out an experiment but observing without influencing the population at all.

Child Development Observation:

Observing children's social interactions on a playground. Researchers might observe children playing in a park and record data on their social behaviors, such as cooperation, conflict resolution, and peer interactions. Observational data collected in natural settings can provide insights into children's social development, peer relationships, and play preferences. This information can be used by educators, psychologists, and parents to support children's social-emotional development.

Analysis of data collected in such ways can broadly categorized into 2 categories called

1. **Descriptive statistics**
2. **Inferential statistics**

1.Descriptive statistics:

Descriptive statistics refers to the branch of statistics that focuses on summarizing and describing the essential features of a dataset.

The characteristics of the data are described in simple terms. Events that are deal with include everyday happenings such as accidents, prices of goods, business, incomes, epidemics, sports data, population data.

Descriptive statistics consists of three basic categories of measures:

1. Measures of central tendency,

2. measures of variability (or spread), and

3. Frequency distribution.

 Measures of central tendency describe the center of the data set (mean,

median, mode).

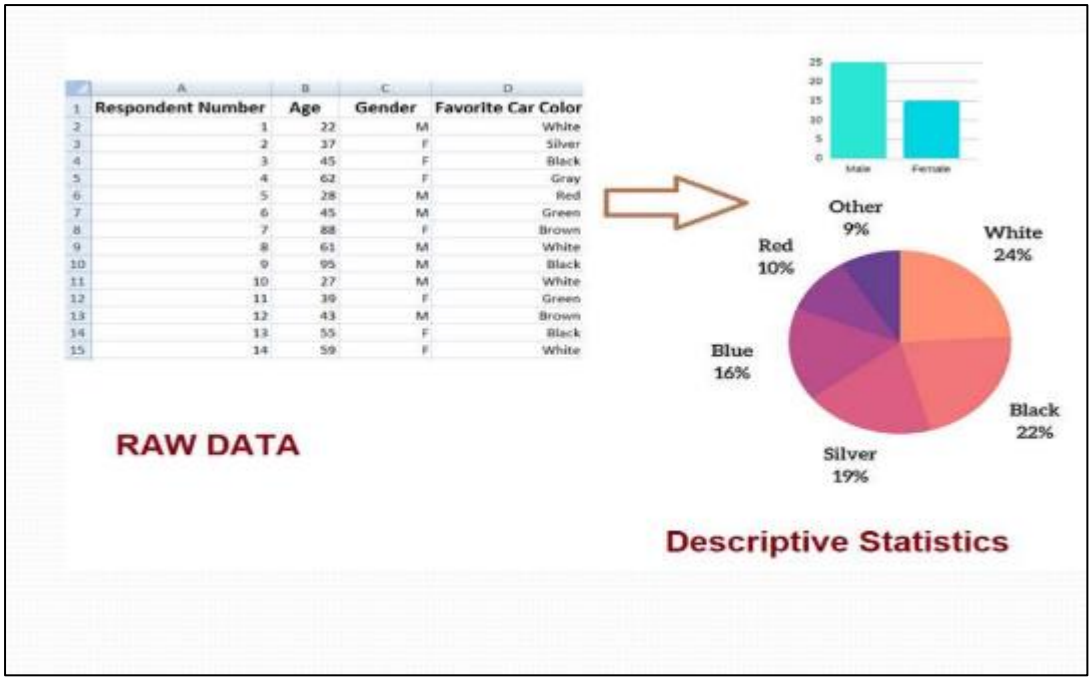
 Measures of variability describe the dispersion of the data set (variance,

standard deviation).

 Measures of frequency distribution describe the occurrence of data within

the data set (count).

Example: Descriptive statistics help you to simplify large amounts of data in a meaningful way. It reduces lots of data into a summary.



- Describe the features of populations and/or samples.
- Organize and present data in a purely factual way.
- Present final results visually, using tables, charts, or graphs.
- Draw conclusions based on known data.
- Use measures like central tendency, distribution, and variance.

Central tendency?

Calculates the middle characteristics of the data (distribution of scores).

 Represent scores in a distribution around which other scores seem to center Most widely

used statistics

Common measures of central tendency include:

 The mean: The average value of all the data points.

 The median: The central or middle value in the dataset.

 The mode: The value that appears most often in the dataset.

Mean: The mean, often referred to as the average, is calculated by summing all the values in a dataset and dividing by the number of values. It's sensitive to extreme values and can be influenced by outliers.

Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santro	19	5

$$\text{Mean (m)} = \frac{\text{Sum of all the terms}}{\text{Total no. of terms}}$$

$$m = \frac{21.3 + 20.8 + 19}{3}$$

$$= 20.366$$

Median: The median is the middle value of a dataset when it's arranged in ascending or descending order. If there is an even number of observations, the median is the average of the two middle values. It's less sensitive to outliers compared to the mean.

Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santro	19	5
i 20	15	4

Ordering the set from lowest to highest = 15 19 20.8 21.3

$$\text{Median} = \frac{19 + 20.8}{2}$$

Median = 23.5

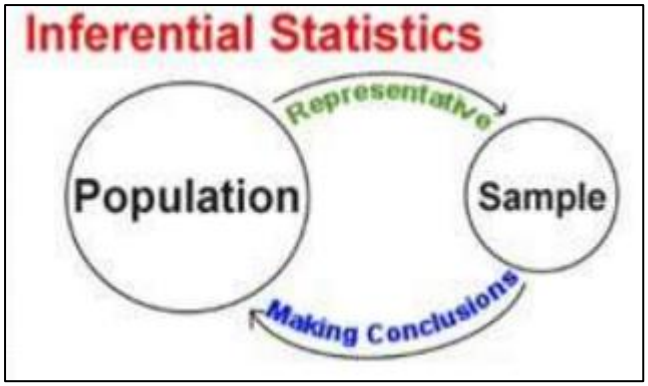
Mode: The mode is the value that appears most frequently in a dataset. A dataset can have one mode (unimodal), two modes (bimodal), or more than two modes (multimodal). Unlike the mean and median, the mode can be used with categorical data as well as numerical data.

2 3 4 2 4 6 4 7 7 4 2 4

Mode = 4

Inferential statistics:

Inferential statistics is a branch of statistics that involves making inferences or predictions about a population based on a sample of data drawn from that population. It's used to draw conclusions, make predictions, or test hypotheses about a larger group (population) based on observations or data collected from a smaller subset (sample) of that population.



Use samples to make generalizations about larger populations.

 Help us to make estimates and predict future outcomes.

 Present final results in the form of probabilities.

 Draw conclusions that go beyond the available data.

 Use techniques like hypothesis testing, confidence intervals, and regression and correlation analysis.

Inferential statistics

 The objective of making inference from data is to make intelligent assertion like

 1. People who don’t smoke live longer than people who smoke.

 2. 80% of all vehicle in USA are 4 wheelers

Random variable :

Whose value cannot be determined before an event happens.

Example:

 1. A person’s blood type.

 2. Number of leaves on a tree

Probability Distribution

• probabilities of occurrence of different possible outcomes for an

experiment.

• Probability distributions are functions that calculates the probabilities of

the outcomes of random variables.

• Typical examples of random variables are:

• coin tosses and dice rolls

Here are some key concepts related to probability distributions:

Random Variable: A random variable is a variable whose possible values are outcomes of a random phenomenon. It can be either discrete (taking on a countable number of distinct values) or continuous (taking on an infinite number of values within a range).

Probability Mass Function (PMF): For discrete random variables, the probability mass function (PMF) gives the probability that the random variable takes on a specific value. It assigns a probability to each possible value of the random variable.

Probability Density Function (PDF): For continuous random variables, the probability density function (PDF) specifies the relative likelihood of different outcomes. Unlike the PMF, the PDF does not directly give probabilities but rather measures the probability density at each point within the range of the random variable.

Cumulative Distribution Function (CDF): The cumulative distribution function (CDF) gives the probability that a random variable takes on a value less than or equal to a given value. It provides a cumulative

view of the distribution, summing up probabilities as we move along the range of the random variable.

Types of Distributions: There are many different probability distributions, each with its own characteristics and applications. Some common ones include:

1. Discrete Distributions: Bernoulli distribution, binomial distribution, Poisson distribution.
2. Continuous Distributions: Normal (Gaussian) distribution, exponential distribution, uniform distribution.
3. Multivariate Distributions: Joint distributions for multiple random variables, such as the multivariate normal distribution.

Properties of Distributions: Probability distributions can vary in terms of their mean, variance, skewness, and kurtosis, which describe the central tendency, spread, symmetry, and shape of the distribution, respectively.

Normal distribution:
The normal distribution, also known as the Gaussian distribution, is one of the most important and widely used probability distributions in statistics. It is characterized by its bell-shaped curve when graphed, with the highest point at the mean and symmetric tails extending indefinitely in both directions.

Here are some key features and properties of the normal distribution:

Symmetry: The normal distribution is symmetric around its mean. This means that the probability of observing a value to the left of the mean is the same as the probability of observing a value to the right of the mean.

Mean, Median, and Mode: In a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution.

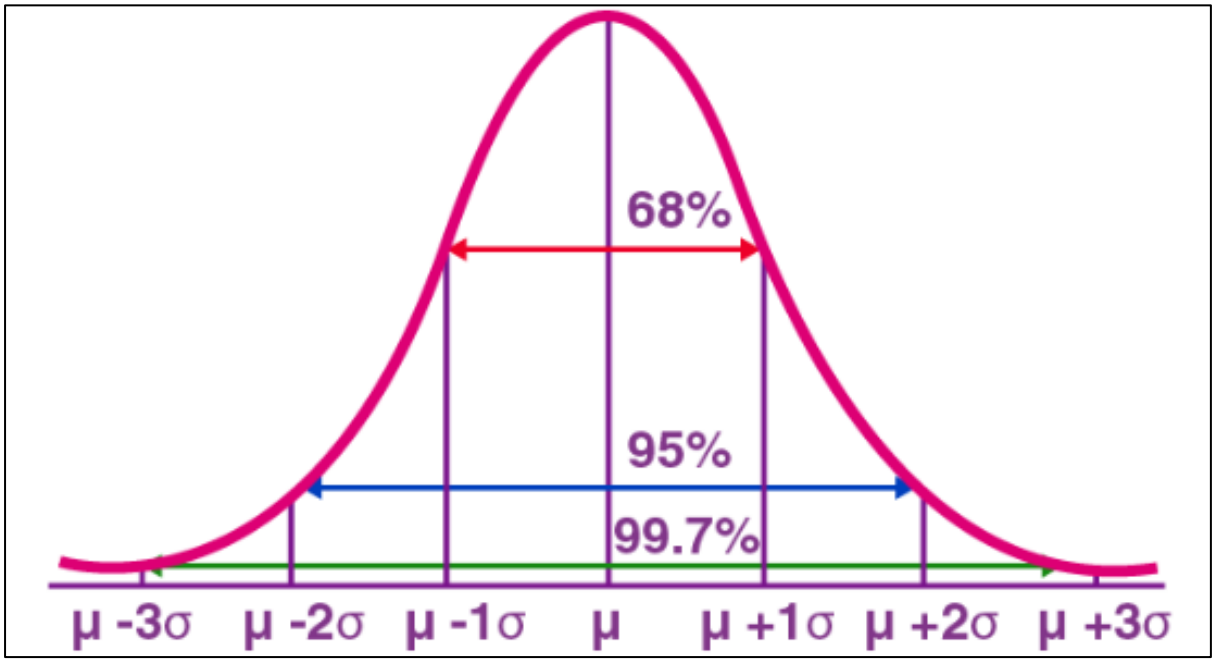
Standard Deviation: The spread or dispersion of the data in a normal distribution is characterized by the standard deviation. About 68% of the data falls within one standard deviation of the mean, 95% falls within two standard deviations, and approximately 99.7% falls within three standard deviations.

Empirical Rule: The empirical rule, also known as the 68-95-99.7 rule, states that in a normal distribution:

Approximately 68% of the data falls within one standard deviation of the mean.

Approximately 95% of the data falls within two standard deviations of the mean.

Approximately 99.7% of the data falls within three standard deviations of the mean.



Measure of Variability

- The goal for variability is to obtain a measure of how spread out the scores are in a distribution.

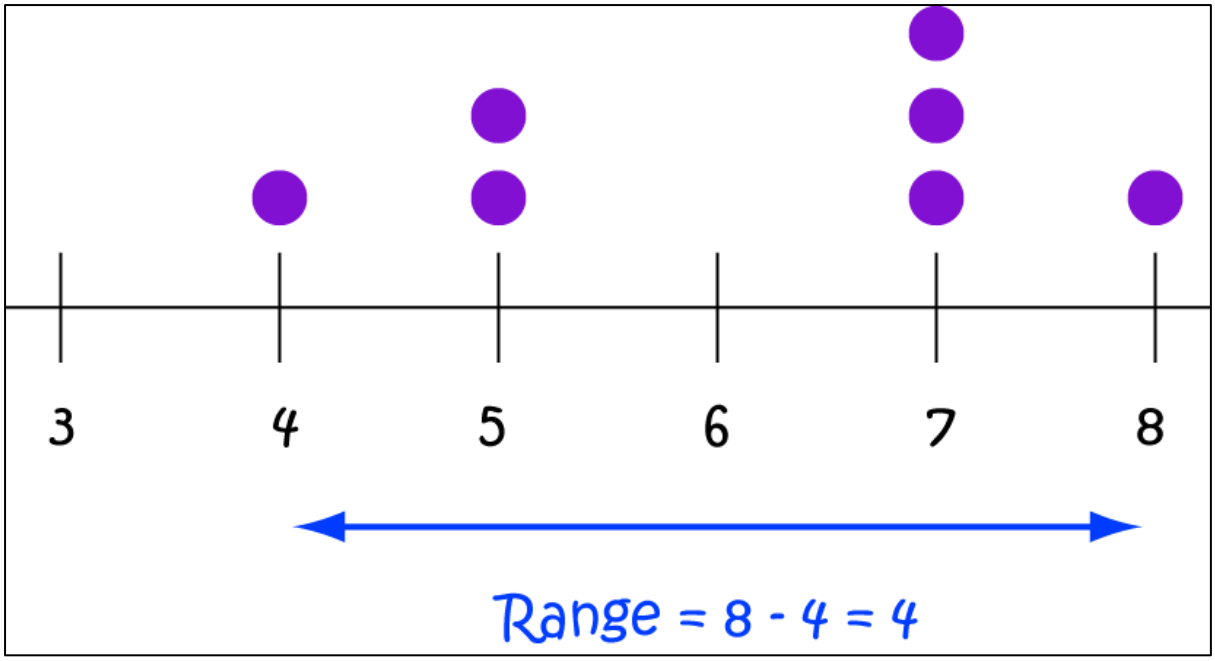
- A measure of variability usually accompanies a measure of central tendency as basic descriptive statistics for a set of scores.
- Central tendency describes the central point of the distribution, and variability describes how the scores are scattered around that central point.
- Together, central tendency and variability are the two primary values that are used to describe a distribution of scores

Measuring Variability

- Variability can be measured with
- the Range
- the Interquartile range
- the Standard Deviation/variance.
- In each case, variability is determined by measuring distance.

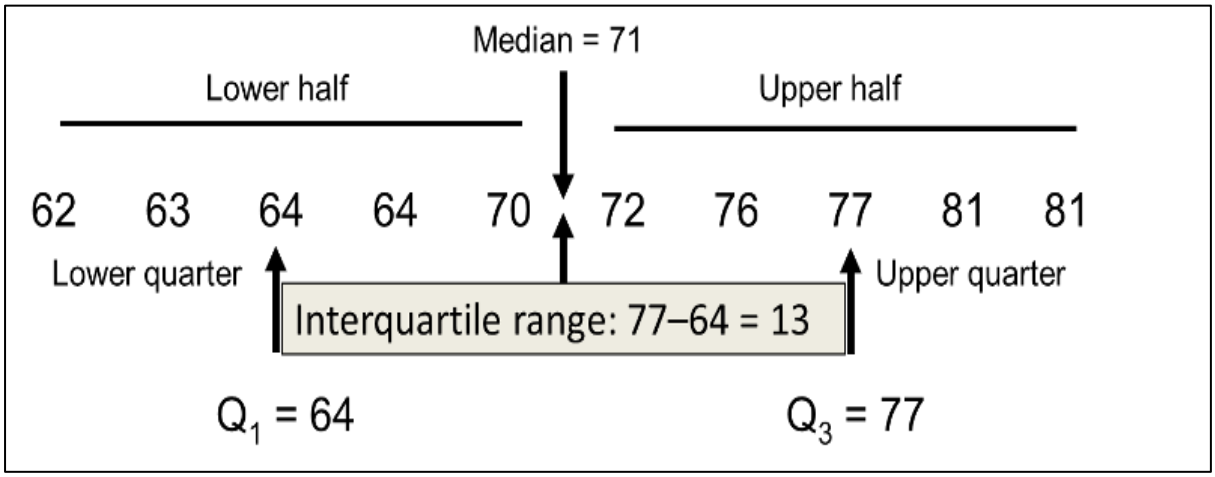
The Range

- The range is the total distance covered by the distribution, from the highest score to the lowest score (using the upper and lower real limits of the range).
- Range tells how wide the data has been distributed



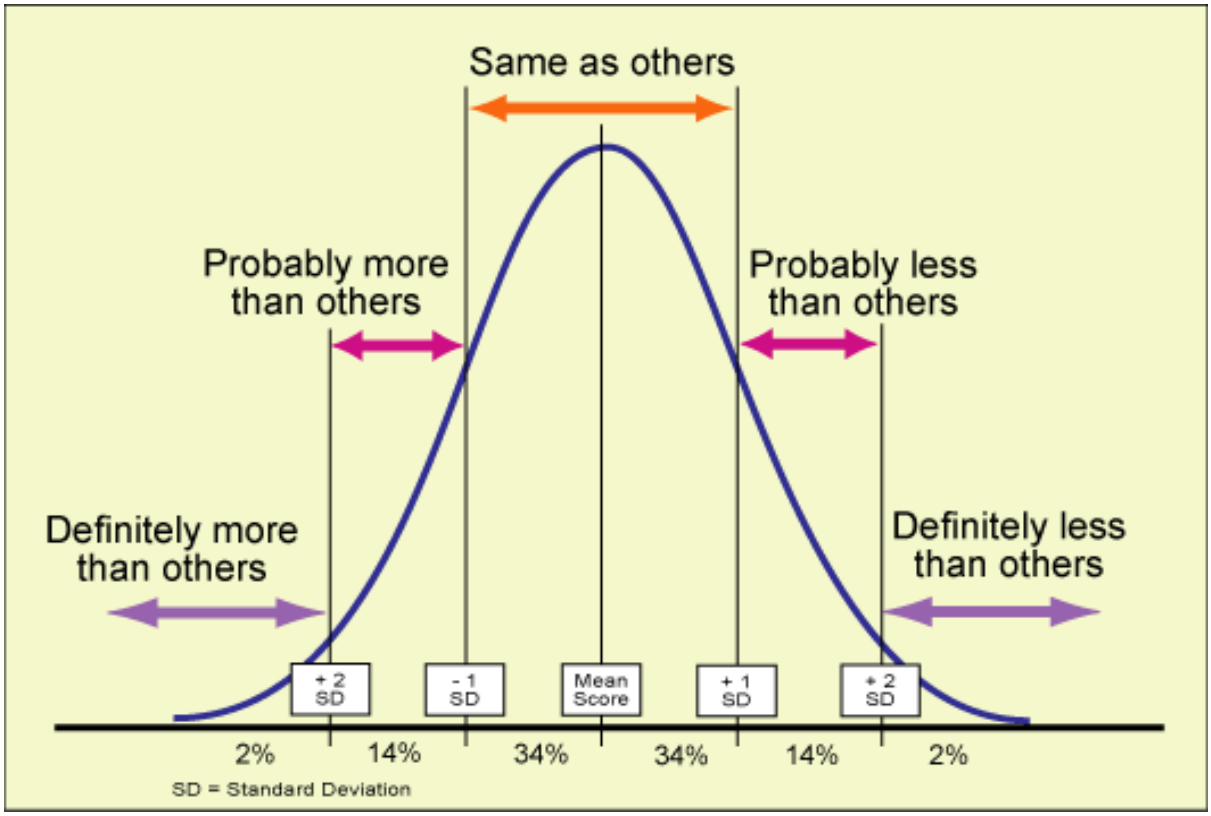
The Interquartile Range:

- The interquartile range is the distance covered by the middle 50% of the distribution.

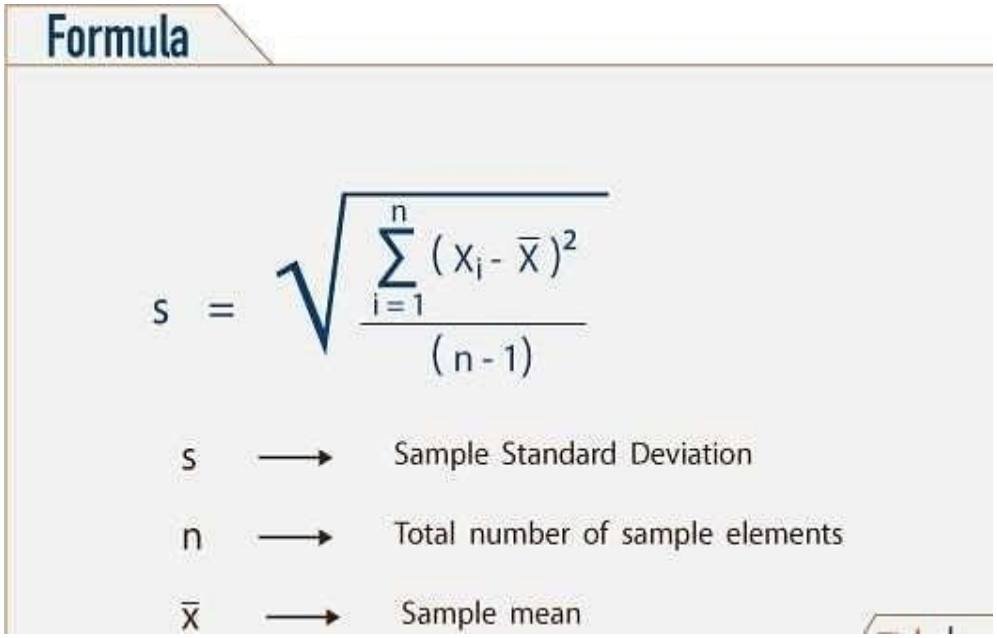


The Standard Deviation:

Standard deviation measures the standard distance between a score and the mean.



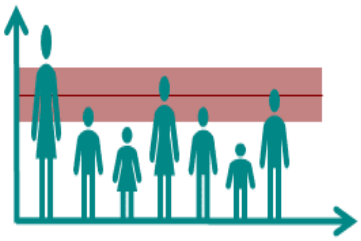
- The calculation of standard deviation can be summarized as a four-step process:
- Compute the deviation (distance from the mean) for each score. Square each deviation.
- Compute the mean of the squared deviations.
- For a population, this involves summing the squared deviations (sum of squares, SS) and then dividing by N.
- The resulting value is called the variance or mean square and measures the average squared distance from the mean.
- Finally, take the square root of the variance to obtain the standard deviation.



Properties of the Standard Deviation

- If a constant is added to every score in a distribution, the standard deviation will not be changed.
- If you visualize the scores in a frequency distribution histogram, then adding a constant will move each score so that the entire distribution is shifted to a new location.
- The center of the distribution (the mean) changes, but the standard deviation remains the same.
- If each score is multiplied by a constant, the standard deviation will be multiplied by the same constant.
- Multiplying by a constant will multiply the distance between scores, and because the standard deviation is a measure of distance, it will also be multiplied

Standard deviation / variance



Average distance of all measured values from the mean value

Range

Distance between lowest and highest value of a distribution

Quantile distance

Spectrum in which the middle 50% of the values lie.

Difference between the first and the third quartile

Example sums : To calculate Standard Deviation

For example: Take the values 2, 1, 3, 2 and 4.

1. Determine the mean (average): 2 + 1 +3 + 2 + 4 = 12

12 ÷ 5 = 2.4 (mean)

Subtract the mean from each value:

2 - 2.4 = -0.4

1 - 2.4 = -1.4

3 - 2.4 = 0.6

2 - 2.4 = -0.4

4 - 2.4 = 1.6

Square each of those differences:

-0.4 x -0.4 = 0.16

-1.4 x -1.4 = 1.96

0.6 x 0.6 = 0.36

-0.4 x -0.4 = 0.16

1.6 x 1.6 = 2.56

Determine the average of those squared numbers to get the variance.

0.16 + 1.96 + 0.36 + 0.16 + 2.56 = 5.2

5.2 ÷ 5 = 1.04 (variance)

Find the square root of the variance.

Square root of 1.04 = 1.01The standard deviation of the values 2, 1, 3, 2 and 4 is 1.01.

Types of sampling:

Simple Random Sampling: This is the most basic form of sampling where each member of the population has an equal chance of being selected, and each selection is independent of the other selections.

Example: Suppose you want to conduct a survey on customer satisfaction at a shopping mall. You could assign each shopper a number and then use a random number generator to select shoppers for the survey.

Stratified Sampling: In stratified sampling, the population is divided into subgroups or strata based on certain characteristics that are important to the study. Then, random samples are taken from each stratum. *Example:* If you're studying income levels in a city, you might divide the population into income brackets (e.g., low-income, middle-income, high-income) and then randomly select individuals from each bracket for your survey.

Cluster Sampling: Cluster sampling involves dividing the population into clusters or groups, usually based on geographical location, and

then randomly selecting entire clusters to be included in the sample. *Example:* If you're studying the effectiveness of a vaccination program in a country, you might randomly select a few cities or districts and then survey all individuals within those selected areas

Systematic Sampling: Systematic sampling involves selecting every nth member from the population, where n is determined by dividing the population size by the desired sample size. *Example:* If you're conducting a study on customer preferences at a grocery store, you might choose to survey every 10th customer who enters the store after starting with a random selection of the first customer.

Sampling Distribution:

A sampling distribution in statistics refers to the distribution of a statistic (such as the mean, variance, proportion, etc.) calculated from many different samples of the same size taken from a population.

Example:

Imagine you want to know the average age of students in a school. The population consists of all students in the school. However, it's not practical to measure the age of every single student due to time and resource constraints.

So, you decide to take several random samples of, let's say, 30 students each from the school population. For each sample, you calculate the mean age of the students in that sample.

After collecting data from multiple samples, you plot a histogram of the sample means. This histogram represents the sampling distribution of the mean age of students in the school.