DAY 1

STATISTICS

WHAT ARE STATISTICS?

Statistics is a branch of mathematics that involves collecting, analysing, interpreting, and presenting data. It provides tools and methods to understand and make sense of large amounts of data and to draw conclusions and make decisions based on the data. In practice, statistics is used in a wide range of fields, such as business, economics, social sciences, medicine, and engineering. It is used to conduct research studies, analyse market trends, evaluate the effectiveness of treatments and interventions, and make forecasts and predictions.

Examples:

1. Business - Data Analysis (Identifying customer behaviour) and Demand Forecasting
2. Medical - Identify efficacy of new medicines (Clinical trials), Identifying risk factor for diseases (Epidemiology)
3. Government & Politics - Conducting surveys, Polling
4. Environmental Science - Climate research

TYPES OF SATISTICS

There are two types of Statistics

1. Descriptive
2. Inferential

Descriptive: - Descriptive statistics deals with the collection, organization, analysis, interpretation, and presentation of data. It focuses on summarizing and describing the main features of a set of data, without making inferences or predictions about the larger population.

Inferential: - Inferential statistics deals with making conclusions and predictions about a population based on a sample. It involves the use of probability theory to estimate the likelihood of certain events occurring, hypothesis testing to determine if a certain claim about a population is supported by the data, and regression analysis to examine the relationships between variables

POPULATION VS SAMPLE

Population refers to the entire group of individuals or objects that we are interested in studying. It is the complete set of observations that we want to make inferences about. For example, the population might be all the students in a particular school or all the cars in a particular city.

A sample, on the other hand, is a subset of the population. It is a smaller group of individuals or objects that we select from the population to study. Samples are used to estimate characteristics of the population, such as the mean or the proportion with a certain attribute. For example, we might randomly select 100 students.

Examples

- All cricket fans vs fans who were present in the stadium
- All students' vs who visits college for lectures

Parameter Vs Statistics

A parameter is a characteristic of a population, while a statistic is a characteristic of a sample. Parameters are generally unknown and are estimated using statistics. The goal of statistical inference is to use the information obtained from the sample to make inferences about the population parameters.

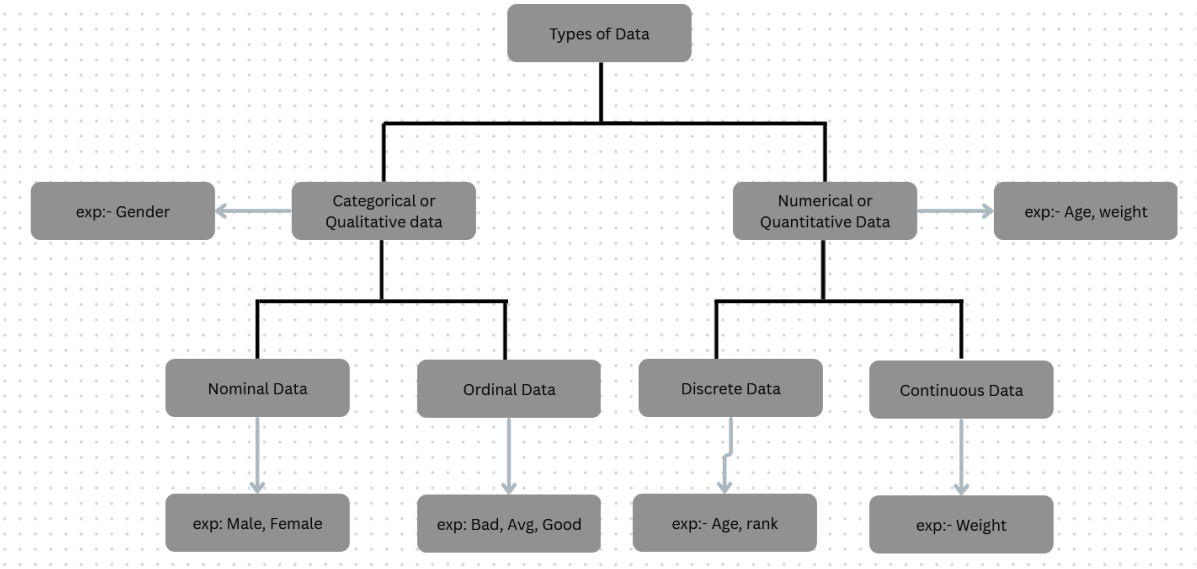
INFERENTIAL STATISTICS

Inferential statistics is a branch of statistics that deals with making inferences or predictions about a larger population based on a sample of data. It involves using statistical techniques to test hypotheses and draw conclusions from data. Some of the topics that come under inferential statistics are:

1. **Hypothesis testing:** This involves testing a hypothesis about a population parameter based on a sample of data. For example, testing whether the mean height of a population is different from a given value.
2. **Confidence intervals:** This involves estimating the range of values that a population parameter could take based on a sample of data. For example, estimating the population mean height within a given confidence level.
3. **Analysis of variance (ANOVA):** This involves comparing means across multiple groups to determine if there are any significant differences. For example, comparing the mean height of individuals from different regions.

4. **Regression analysis:** This involves modelling the relationship between a dependent variable and one or more independent variables. For example, predicting the sales of a product based on advertising expenditure.
5. **Chi-square tests:** This involves testing the independence or association between two categorical variables. For example, testing whether gender and occupation are independent variables.
6. **Sampling techniques:** This involves ensuring that the sample of data is representative of the population. For example, using random sampling to select individuals from a population.
7. **Bayesian statistics:** This is an alternative approach to statistical inference that involves updating beliefs about the probability of an event based on new evidence. For example, updating the probability of a disease given a positive test result.

TYPES OF DATA



MEASURE OF CENTRAL TENDENCY

A measure of central tendency is a statistical measure that represents a typical or central value for a dataset. It provides a summary of the data by identifying a single value that is most representative of the dataset as a whole.

There are three types:

1. **Mean**
2. **Median**
3. **Mode**

Mean: The mean is the sum of all values in the dataset divided by the number of values.

Exp: 3,4,1,2,5 = 15/5 = 3

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

Drawback: It influences by Outliers

Median: The median is the middle value in the dataset when the data is arranged in order.

Exp:

2,3,1,2,5 = arrange in order = 1,2,3,4,5 = odd = we choose middle one = 3

2,3,1,2,5,600 = arrange in order = 1,2,3,4,5,600 = even = we add two middle numbers then find the mean = (3+4)/2 = 3.5

Note: It does not influence by Outliers.

Mode: - The mode is the value that appears most frequently in the dataset.

Example : - 1,3,4,6,1,2,3,5,1,2

Ans : 1

Weighted Mean: - The weighted mean is the sum of the products of each value and its weight, divided by the sum of the weights. It is used to calculate a mean when the values in the dataset have different importance or frequency.

Trimmed Mean: - A trimmed mean is calculated by removing a certain percentage of the smallest and largest values from the dataset and then taking the mean of the remaining values. The percentage of values removed is called the trimming percentage.

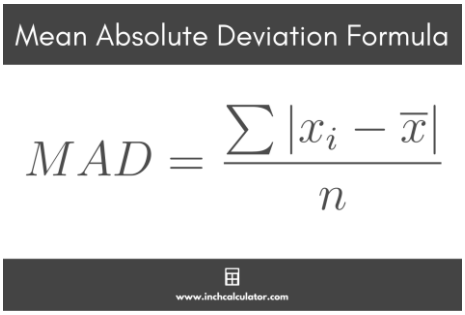
MEASURE OF DISPERSION

A measure of dispersion is a statistical measure that describes the spread or variability of a dataset. It provides information about how the data is distributed around the central tendency (mean, median or mode) of the dataset.

1. **Range:** - The range is the difference between the maximum and minimum values in the dataset. It is a simple measure of dispersion that is easy to calculate but can be affected by outliers.
2. **Variance:** - The variance is the average of the squared differences between each data point and the mean. It measures the average distance of each data point from the mean and is useful in comparing the dispersion of datasets with different means.

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element $\bar{x}$ = sample mean n = sample size

Mean absolute variance



3. **Standard Deviation: -** The standard deviation is the square root of the variance. It is a widely used measure of dispersion that is useful in describing the shape of a distribution.

Population	Sample
$\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2}{n}}$	$S = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}$
μ - Population Average xi - Individual Population Value n - Total Number of Population	x - Sample Average x_i - Individual Population Value n - Total Number of Sample

4. **Coefficient of Variation:** -The CV is the ratio of the standard deviation to the mean expressed as a percentage. It is used to compare the variability of datasets with different means and is commonly used in fields such as biology, chemistry, and engineering.

   The coefficient of variation (CV) is a statistical measure that expresses the amount of variability in a dataset relative to the mean. It is a dimensionless quantity that is expressed as a percentage.

   The formula for calculating the coefficient of variation is:

   CV = (standard deviation / mean) x 100%

   GRAPS FOR UNIVARIATE ANLYSIS

   There are two types of:

1. Categorical
2. Numerical

Categorical: - **Frequency Distribution Table & Cumulative Frequency**

A frequency distribution table is a table that summarizes the number of times (or frequency) that each value occurs in a dataset.

Let's say we have a survey of 200 people, and we ask them about their favourite type of vacation, which could be one of six categories: Beach, City, Adventure, Nature, Cruise, or Other

Type of Vacation	Frequency
Beach	60
City	40
Adventure	30
Nature	35
Cruise	20
Other	15

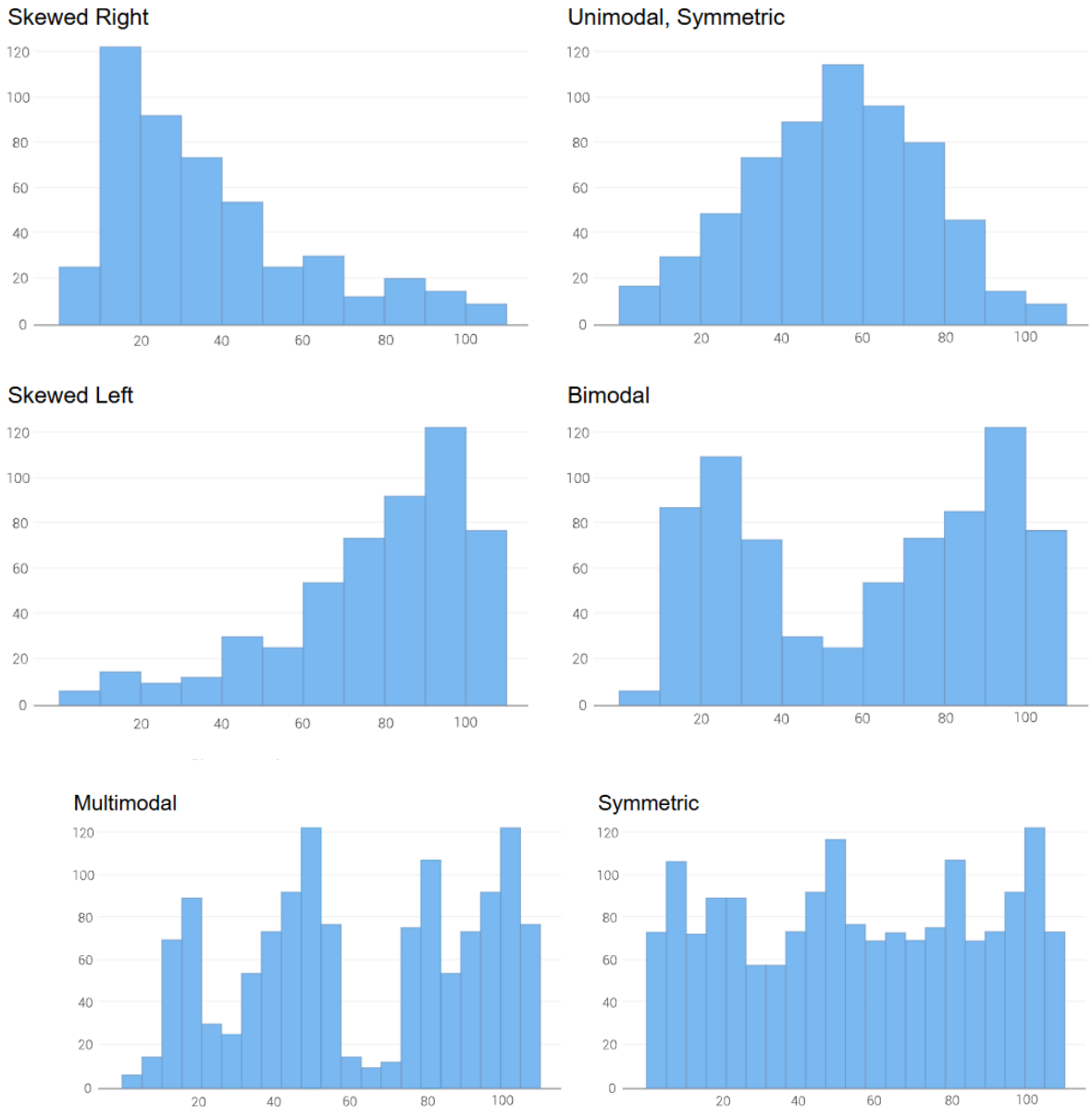
Relative frequency is the proportion or percentage of a category in a dataset or sample. It is calculated by dividing the frequency of a category by the total number of observations in the dataset or sample.

Type of Vacation	Frequency	Relative Frequency
Beach	60	0.3
City	40	0.2
Adventure	30	0.15
Nature	35	0.175
Cruise	20	0.1
Other	15	0.075

Cumulative frequency is the running total of frequencies of a variable or category in a dataset or sample. It is calculated by adding up the frequencies of the current category and all previous categories in the dataset or sample.

Type of Vacation	Frequency	Relative Frequency	Cumulative Frequency
Beach	60	0.3	60
City	40	0.2	100
Adventure	30	0.15	130
Nature	35	0.175	165
Cruise	20	0.1	185
Other	15	0.075	200

Numerical - Frequency Distribution Table & Histogram



GRAPHS FOR BIVARIATE ANALYSIS

There are three types of:

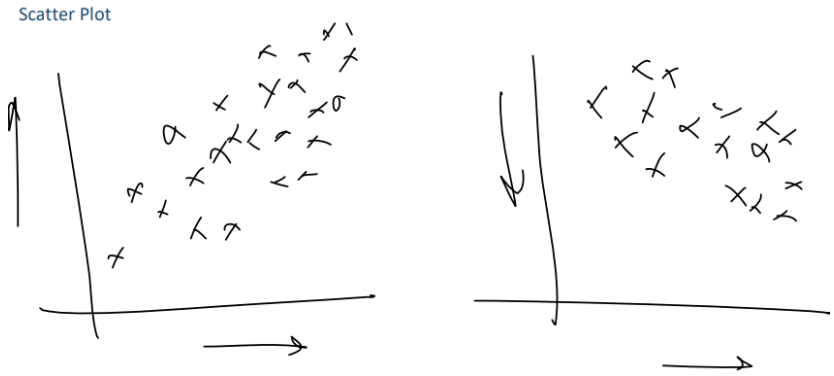
1. Categorical – Categorical
2. Numerical – Numerical
3. Categorical – Numerical

Categorical – Categorical: -

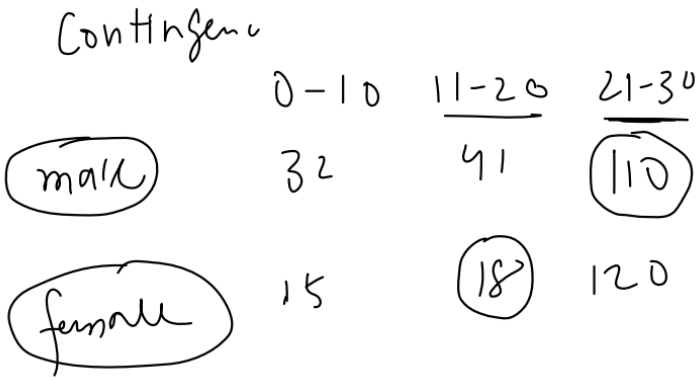
Contingency Table/Crosstab

A contingency table, also known as a cross-tabulation or crosstab, is a type of table used in statistics to summarize the relationship between two categorical variables. A contingency table displays the frequencies or relative frequencies of the observed values of the two variables, organized into rows and columns.

Numerical – Numerical: -



Categorical – Numerical:-



DATA BASE MANGEMENT SYSTEM(DBMS)

WHAT ARE DATBASES?

A Database is a shared collection of logically related data and description of these data, designed to meet the information needs of an organization.

Data Storage: A database is used to store large amounts of structured data, making it easily accessible, searchable, and retrievable.

Data Analysis: A database can be used to perform complex data analysis, generate reports, and provide insights into the data.

Record Keeping: A database is often used to keep track of important records, such as financial transactions, customer information, and inventory levels.

Web Applications: Databases are an essential component of many web applications, providing dynamic content and user management.

PROPERTIES OD AN IDEAL DATABASE

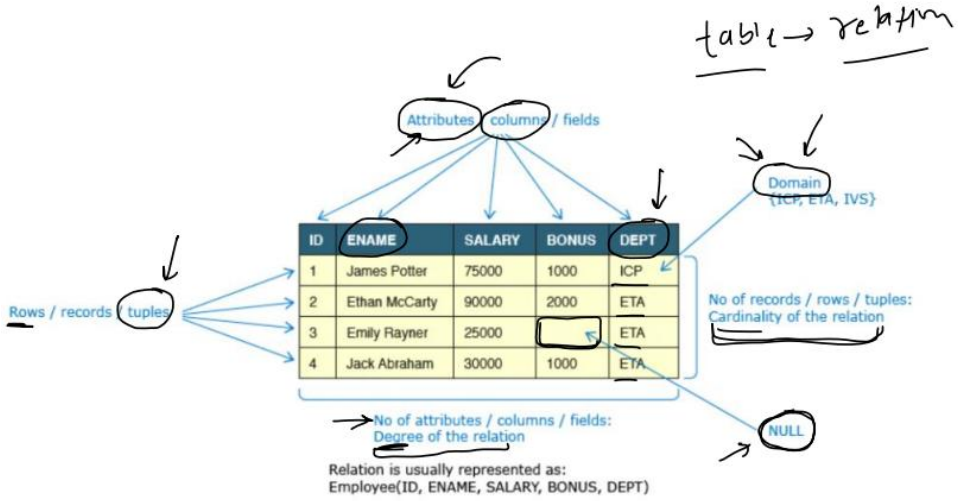
- Integrity
- Availability
- Security
- Independent of Application
- Concurrency

TYPES OF DATABASES

1. **Relational Databases -** Also known as SQL databases, these databases use a relational model to organize data into tables with rows and columns.
2. **NoSQL Databases -** These databases are designed to handle large amounts of unstructured or semi-structured data, such as documents, images, or videos. (MongoDB)
3. **Column Databases** - These databases store data in columns rather than rows, making them well suited for data warehousing and analytical applications. (Amazon Redshift, Google Big Query)
4. **Graph Databases -** These databases are used to store and query graph-structured data, such as social network connections or recommendation systems. (Neo4j, Amazon Neptune)
5. **Graph Databases** - These databases are used to store and query graph-structured data, such as social network connections or recommendation systems. (Neo4j, Amazon Neptune)

RELATIONAL DATABASE

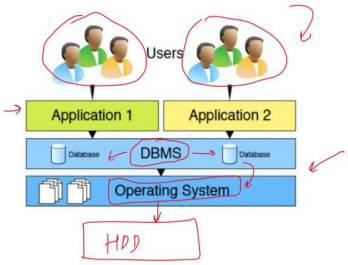
Also known as SQL databases, these databases use a relational model to organize data into tables with rows and columns.



ID	ENAME	SALARY	BONUS	DEPT
1	James Potter	75000	1000	ICP
2	Ethan McCarty	90000	2000	ETA
3	Emily Rayner	25000		ETA
4	Jack Abraham	30000	1000	ETA

WHAT IS DBMS

A database management system (DBMS) is a software system that provides the interfaces and tools needed to store, organize, and manage data in a database. A DBMS acts as an intermediary between the database and the applications or users that access the data stored in the database.



CORE FUNCTION ALITIES OF DBMS

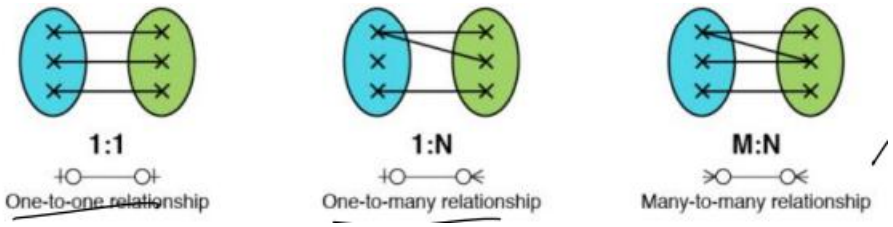
- **Data Management** - Store, retrieve and modify data
- **Integrity** - Maintain accuracy of data
- **Concurrency** - Simultaneous data access for multiple users
- **Transaction** - Modification to database must either be successful or must not happen at all
- **Security** - Access to authorized users only
- **Utilities** - Data import/export, user management, backup, logging

DATABASE KEYS

- **Super Key** - A Super key is a combination of columns that uniquely identifies any row within a relational database management system (RDBMS) table
- **Candidate key** - A candidate key is a minimal Super key, meaning it has no redundant attributes. In other words, it's the smallest set of attributes that can be used to uniquely identify a tuple (row) in the table
- **Primary Key** - A primary key is a unique identifier for each tuple in a table. There can only be one primary key in a table, and it cannot contain null values.
- **Alternate Key** - An alternate key is a candidate key that is not used as the primary key.
- **Composite Key** - A composite key is a primary key that is made up of two or more attributes. Composite keys are used when a single attribute is not sufficient to uniquely identify a tuple in a table
- **Surrogate Key -** A surrogate key is a unique identifier for a row in a database table that is not derived from the data itself. It's often an automatically generated integer, and it serves as a primary key when a natural key (a key based on existing data attributes) is not suitable or available.
- **Foreign Key -** A foreign key is a primary key from one table that is used to establish a relationship with another table.

CARDINALITY OF RELATIONSHIP

Cardinality in database relationships refers to the number of occurrences of an entity in a relationship with another entity. Cardinality defines the number of instances of one entity that can be associated with a single instance of the related entity.



Examples

- Person -> Driving License Number
- Student -> college branch
- Restaurants -> orders
- Restaurants -> menu
- Students -> courses