STATISTICS

We will discuss these below usecases after studying Statistics:

USECASE 1: (Asked in a Product Based Company)

Bank: HDFC

HDFC have already two ATM in Location A and B. Distance between location A and B is 30km. They are planning to open a new ATM C, Whether to open ATM in C location or not?

This task is on Data Analyst and Data Scientist.

USECASE 2: (Asked in Amazon for Data Analyst Position)

Find the average size of the shark throughout the world?

USECASE 3: (Asked in Intuit Product Based Company)

Which day of the month should be a Big Billion Day?

LIFE CYCLE OF DATA SCIENCE PROJECT:

Requirement Gathering

Business Analyst, Product Manager, Project Manager will be involved

Data Analytics

Requirements will be provided to data team-->Data Analyst, Data Scientist, Big Data Enggs, Cloud Enggs. To solve the problem, where should we get the data? Data Analyst, Data Scientist, Business Analyst, Product Managers will be involved in this process. They will have domain/business knowledge understanding. This data could be internal data, 3rd party API's, Live Streaming etc. Once we get all the data, data will be sent to Big Data Engg Team. They will save the data in databases (MySQL, NoSQL)

DATA SCIENCE PROJECT STEPS:

- EDA (Exploratory Data Analysis includes Cleaning of Data)

  In statistics, exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.

- Feature Engineering (E.g. Handling Missing Values)
  - Feature engineering or feature extraction or feature discovery is the process of using domain knowledge to extract features from raw data.
  - Feature Engineering encapsulates various data engineering techniques such as selecting relevant features, **handling missing data, encoding the data, and normalizing** it. It is one of the most crucial tasks and plays a major role in determining the outcome of a model.
- Feature Selection
  - In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction.

- Feature selection is the process by which a subset of relevant features, or variables, are selected from a larger data set for constructing models. Variable selection, attribute selection or variable subset selection are all other names used for feature selection.

- Model Training

  Model training is the phase in the data science development lifecycle where practitioners try to fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range.

- Hyperparameter Tuning (To improve the performance of the model)

  - Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

- Deploy the Model

Note:
For Feature Engineering, Feature Selection, Model Training, Hyperparameter Tuning Statistics required everywhere. Hence it is important to know statistics.

WHAT IS STATISTICS

Statistics is the science of collecting, organizing and analyzing the data.

What is Data?

Data is Facts or Pieces of Information.

e.g. Salary of the employees, Designation of the Employees

TYPES OF STATISTICS:

DESCRIPTIVE STATISTICS:

It Consist of Organizing and Summarizing the Data using different kind of plots:

Mean-Median-Mode, Standard Deviation, Variance etc.

Measure of Central Tendency (Eg. Plotting Using Pie-Charts, Histogram, Bar Chart,

Distribution (We draw using Histogram, Candlestick, Box Plots, Whisker Plot, Skewed Plots, Scatter Plots, Violin Plots). This will be used extensively in EDA and Feature Engineering.

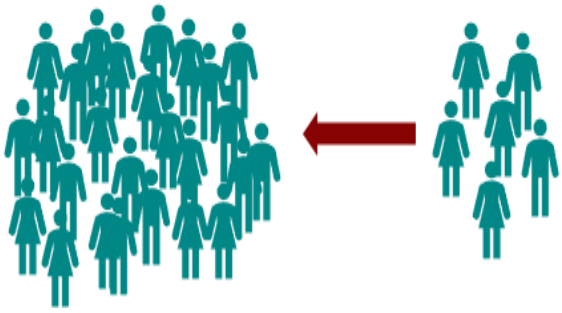
INFERENTIAL STATISTICS:

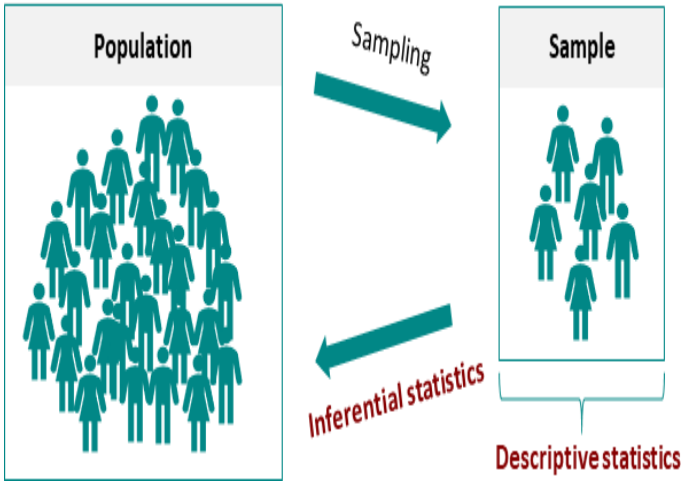
It Consist of Collecting Sample Data and **making conclusions** (Using Hypothesis Testing) about Population Data using some experiments.

Let's say there is a university with 500 students. Class A has 60 people. By using Age data of these 60 people (sample), can we make conclusions of Average age of entire university of 500 students ? (YES) We will be using Technique of Hyothesis Testing, Confidence Interval, P Value, Experiments like Z Test, T Test, Chi Square Test, Anova or F Test.

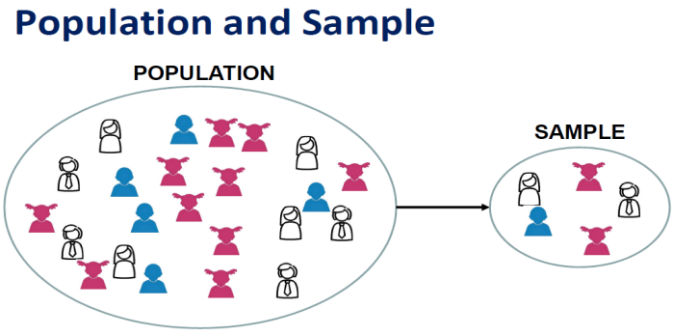
Inferential statistics:

Testing of statements about the population on the basis of sample characteristics.





Sample Data vs Population Data:



E.g. -Punjab Population 10 Crore

Exit Polls conducted: Asking people whom did you vote and make conclusions. They don't ask each and every people as it is not practically possible. So they need to take sample of data.

E.g. -Lets say there are 20 Classrooms in a University and you have collected Ages of students in 1 Classroom:

Ages: {21,20,18,34,17,22,24,25} Weight: {.....}

What Descriptive Stats Qns can come:

What is the Average Age for that particular Class?

Relationship between Age and Weight?

What Descriptive Inferential Stats Qns can come:

Are the average age of the students in the classroom less than the average age of the students in the university?

1000 Students->Class A 50 girls (95% Avg Marks) and Class B 50 boys (92% Avg Marks). Have girls performed better than boys in entire University ?

DIFFERENT SAMPLING TECHNIQUES:

- **Simple Random Sampling:** Every member of the population (N) has an equal chance of being selected for your sample (n)
  E.g. Exit Polls, General Survey, Lottery

- **Stratified Sampling:** Strata means Layers/Clusters/Groups: E.g. We can make groups basis Gender, Education Degree etc.

- **Systematic Sampling**: (In Airport, Credit Card Section wants to sell credit card: Suppose Two person are selling credit card, 1st individual said Every 5th person I will see I will approach for credit card, 2nd individual said Every 9th person I will see I will approach for credit card).
  Basically selecting every nth individual out of Population (N)
- **Convenience Sampling:** Only those who are interested in the survey will only participate.
  E.g. Data Science Survey

E.g-Survey Regarding New Technology - Convenience Sampling (bcz those who are interested in technology)

E.g.-RBI Survey:(specific survey for women who are taking care of entire house)- Stratified sampling(bcz survey for married women) then random sampling

E.g-Selling Credit Cards : We can apply Stratified Sampling (credit card directly based on salary) first and then on that sample, we can apply Random Sampling

Variable:

Variable is a Property that can take any values E.g. Age = 14

Variables will have multiple values (Collections): Ages = [24,25,26,27]

Two Different Type of Variables:

- **Quantitative Variable**: Measured Numerically. We can do mathematical operations: Eg. Age, Weight, Height, Rainfall in cm This can be further divided into Discrete and Continuous Variables:
  - Discrete (Cannot be Decimal): Number of Bank Account Person can have, No. of children in family
  - Continuous (Decimal allowed): Height, weight, ages, rainfall, temperature, distance
- **Qualitative Variable**: Categorical Variables: E.g. Based on Some Characteristics, they are grouped together.

What kind of Variable is :

- Marital Status? -Categorical Variable
- Ganga River Length? -Continuous
- Movie Duration? -Continuous
- Pin code? -Discrete
- IQ? - Discrete
- Gender? -Categorical