prediction

analysis, pattern finding / modelling

collection, organization, processing,

Everything about data

Data Science

tools, processes, roles

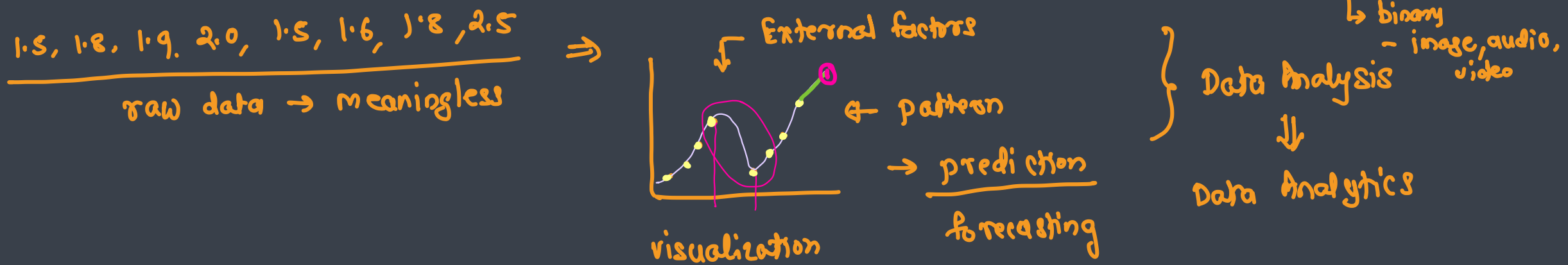
Data Science = maths/stats + Domain knowledge (SME) + AI/ML/DL/GenAI

Data Analysis + Data Analytics + programmer + data

past future

What is Data Science

- Data science is the domain of study that deals with vast volumes of data using modern tools and techniques, including essential data science skills, to find unseen patterns, derive meaningful information, and make business decisions
- Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI) and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data. These insights can be used to guide decision making and strategic planning
- Data science uses complex machine learning algorithms to build predictive models
- The data used for analysis can come from many different sources and presented in various formats



Data

Structured

→ Database

RDBMS

id	name	email

Semi-structured

→ json, xml, Excel
csv, tsv

```
[
  { "name": "...",
    "email": "..."
  },
  :
]
```

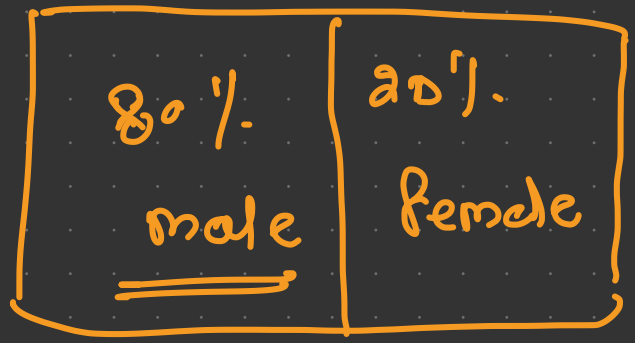
Unstructured

→ image, audio,
video, text

Data wrangling
Data cleansing

- check the missing data
- check the proper data types
- check the data imbalance
- check the scale

age = "30"



age	salary
20-80	10k-10L

Data Science Lifecycle

end-to-end process

- The data science lifecycle involves various roles, tools, and processes, which enables analysts to glean actionable insights
- Data science's lifecycle consists of five distinct stages, each with its own tasks:

1. **Capture** → data collection → primary or secondary
   - Data Acquisition, Data Entry, Signal Reception, Data Extraction
   - This stage involves gathering raw structured and unstructured data
2. **Maintain** → organization of data
   - Data Warehousing, Data Cleansing, Data Staging, Data Processing, Data Architecture
   - This stage covers taking the raw data and putting it in a form that can be used

model = formula

3. **Process** (finding, grouping, formulation, NLP)
   - Data Mining, Clustering/Classification, Data Modeling, Data Summarization (analysis)
   - Data scientists take the prepared data and examine its patterns, ranges, and biases to determine how useful it will be in predictive analysis
4. **Analyze** (EDA, forecast, visualization – charts/tables/graphs)
   - Exploratory/Confirmatory, Predictive Analysis, Regression, Text Mining, Qualitative Analysis
   - Here is the real meat of the lifecycle. This stage involves performing the various analyses on the data

↳ strictly performed by human, → cannot be automated

Data Science Lifecycle

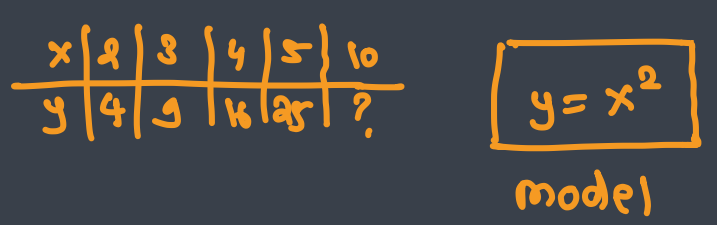
5. **Communicate** → Reporting
   - Data Reporting, Data Visualization, Business Intelligence, Decision Making
   - In this final step, analysts prepare the analyses in easily readable forms such as charts, graphs, and reports

visualization
- → for analysis
- → Reporting

} charts / graphs / tables / figures / dashboards

stackholder
- entity interested and affected by the project
- VC, client, programmers, users

Data Science Prerequisites



- **Machine Learning**
  - Machine learning is the backbone of data science
  - Data Scientists need to have a solid grasp of ML in addition to basic knowledge of statistics
- **Modeling** → formulation analysis
  - Mathematical models enable you to make quick calculations and predictions based on what you already know about the data
  - Modeling is also a part of Machine Learning and involves identifying which algorithm is the most suitable to solve a given problem and how to train these models
- **Statistics**
  - Statistics are at the core of data science
  - A sturdy handle on statistics can help you extract more intelligence and obtain more meaningful results
- **Programming**
  - Some level of programming is required to execute a successful data science project
  - The most common programming languages are Python, and R
  - Python is especially popular because it's easy to learn, and it supports multiple libraries for data science and ML
- **Database**
  - A capable data scientist needs to understand how databases work, how to manage them, and how to extract data from them

Data Science Roles

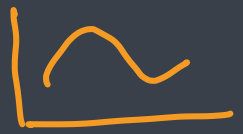
- Data Strategist → data collection → Scraping
  - Ideally, before a company collects any data, it hires a data strategist—a senior professional who understands how data can create value for businesses
  - They can make data-driven decisions
  - Data can help create smarter products and services
  - Companies can use data to improve business processes
  - They could create a new revenue stream via data monetization
  - Companies often outsource such data roles. They hire external consultants to devise a plan that aligns with the organizational strategy
- Data Architect
  - A data architect (or data modeler) plans out high-level database structures
  - This involves the planning, organization, and management of information within a firm, ensuring its accuracy and accessibility. In addition, they must assess the needs of business stakeholders and optimize schemas to address them
  - Without proper data architecture, key business questions may remain answered due to the lack of coherence between different tables in the database
  - A data architect is a senior professional and often a consultant. To become one, you'd need a solid resume and rigorous preparation for the interview process.

Data Science Roles

- **Data Engineer**
  - The role of data engineers and data architects often overlaps—especially in smaller businesse
  - Data engineers build the infrastructure, organize tables, and set up the data to match the use cases defined by the architect
  - They handle the ETL process, which stands for Extract, Transform, and Load
  - This involves retrieving data, processing it in a usable format, and moving it to a repository
  - Simply put, they pipe data into tables correctly

- **Data Analyst**
  - Data analysts explore, clean, analyze, visualize, and present information, providing valuable insights for the business
  - They typically use SQL to access the database
  - They leverage an object-oriented programming language like Python or R to clean and analyze data and rely on visualization tools, such as Power BI or Tableau, to present the findings

Data Science Roles



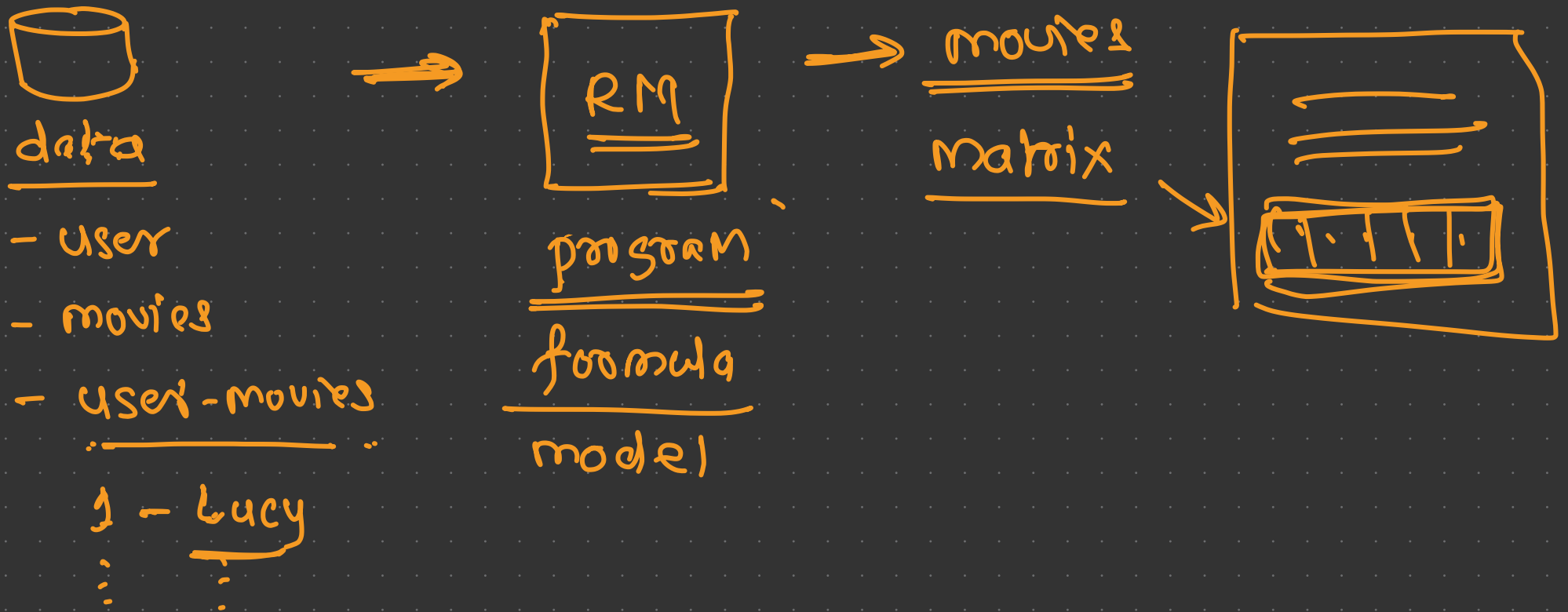
- **Business Intelligence Analyst**
  - Data analyst's and BI analyst's duties overlap to a certain extent, but the latter has more of a reporting role
  - Their main focus is on building meaningful reports and dashboards and updating them frequently.
  - More importantly, they have to satisfy stakeholders' informational needs at different levels of the organization
- **Data Scientist**
  - Data scientist has the skills of a data analyst but can leverage machine learning to create models and make predictions based on past data
  - Three main types of data scientists:
    - Traditional data scientists → Shoots
      - Scientist does all sorts of tasks, including data exploration, advanced statistical modeling, experimentation via A/B testing, and building and tuning machine learning models
    - Research scientists
      - Research scientists primarily work on developing new machine learning models for large companies
    - Applied scientists
      - Frequently hired in big tech and larger companies—boast one of the highest-paid jobs in data science
      - These specialists combine data science and software engineering skills to productionize models

Data Science Roles

- **ML Ops Engineer**
  - Companies that don't have applied scientists hire ML Ops engineers
  - They are responsible for putting the ML models prepared by traditional data scientists into production
  - In many instances, ML Ops engineers are former data scientists who have developed an engineering skillset
  - Their main responsibilities are to put the ML model in production and fix it if something breaks
- **Data Product Manager**
  - The last role we discuss in this article is that of a product manager
  - The person in this position is accountable for the success of a data product
  - They consider the bigger picture, identifying what product needs to be created, when to build it, and what resources are necessary
  - A significant focus of such data science roles is data availability—determining whether to collect data internally or find ways to acquire it externally
  - Ultimately, product managers strategize the most effective ways to execute the production process

Netflix → video streaming → Recommendation Model

ML Engineer



Data Scientist Tasks

- Know enough about the business to ask pertinent questions and identify business pain points
- Apply statistics and computer science, along with business acumen, to data analysis
- Use a wide range of tools and techniques for preparing and extracting data—everything from databases and SQL to data mining to data integration methods
- Extract insights from big data using predictive analytics and artificial intelligence (AI), including machine learning models, natural language processing, and deep learning
- Write programs that automate data processing and calculations
- Tell—and illustrate—stories that clearly convey the meaning of results to decision-makers and stakeholders at every level of technical understanding
- Explain how the results can be used to solve business problems
- Collaborate with other data science team members, such as data and business analysts, IT architects, data engineers, and application developers

Data Science Tools

- **Data Analysis**
  - SAS, Jupyter, RStudio, MATLAB, Excel, RapidMiner
- **Data Warehousing**
  - Informatica/ Talend, AWS Redshift
- **Data Visualization**
  - Jupyter, Tableau, Cognos, RAW
- **Machine Learning**
  - Spark MLib, Mahout, Azure ML studio

programming languages - python / R

python → numpy, pandas, matplotlib, seaborn, selenium, scrapy,

ML - sci-kit, XGBoost, AdaBoost

DL - tensorflow, pytorch

NLP → transformers, NLTK, spacy

third party → Gemini

Data Science Applications

- **Healthcare**
  - Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases
- **Gaming**
  - Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level
- **Image Recognition**
  - Identifying patterns is one of the most commonly known applications of data science
  - In images and detecting objects in an image is one of the most popular data science applications
- **Recommendation Systems**
  - Netflix and Amazon give movie and product recommendations based on what you like to watch, purchase, or browse on their platforms
- **Logistics**
  - It is used by logistics companies to optimize routes to ensure faster delivery of products and increase operational efficiency
- **Fraud Detection**
  - Banking and financial institutions use data science and related algorithms to detect fraudulent transactions

Statistics

Basic knowledge

Introduction

primary secondary → database → File Repository → csv, json → visualization → chart/graph → human

- Discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data → Reporting - charts / graphs / tables
- A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data
- It is a branch of scientific method used in dealing with phenomena that can be described numerically either by counts or by measurements → numerical data
- The word "statistics" refers either to quantitative information or to a method of dealing with quantitative information
- Study of statistics involves methods of refining numerical and non-numerical information into useful forms like charts, tables, figures etc.
- The methods by which statistical data are analysed are called as statistical methods

ML is extension to stats

Regression

What is statistics ?

- Aggregation of facts
- It is numerically expressed
- It is enumerated or estimated according to Reasonable standards of accuracy
- It is collected in systematic manner
- It is collected for predetermined purpose
- It should be placed in relation to each other
- All statistics are numerical statements of facts, but all numerical statements of facts are not statistics

Statistical Investigations

sources → primary, secondary

physical survey
- user name
- age
- address

- **Collection** – capture → Data injestion
  - First stage of statistical investigation
  - Utmost care must be exercised in collecting the data because they form the foundation of analysis
  - If the data is faulty then conclusion drawn will never be reliable
  - The data may be available from existing published or unpublished sources or collected explicitly by researcher

50%

- **Organization** physical to digital format →
  - First step of organization is Editing the data → data cleansing
    - Adjusting data for omissions, inconsistencies, irrelevant answers and wrong computations
  - Second step after editing is Classification
    - Arranging data according to some common characteristics possessed by items constituting the data → categorizing / clustering
  - Last step is tabulation
    - Arranging the data into rows and columns to get more clarity

records

Statistical Investigations

- **Presentation** → visualization for analysis
  - Data presented in an orderly manner facilitate statistical analysis
  - There are many ways the data can be presented such as diagrams, graphs, charts, tables etc.
- **Analysis**
  - The purpose of analysis of data is to dig out information useful for decision making
  - Methods used for analysis
    - Measures of central tendency : mean, mode and median
    - Measures of variations : range, quartiles, IQR, variance, standard deviation
    - Measures of Skewness : skewness and kurtosis
    - Measures of relationships : covariance, correlation and regression
- **Interpretation** - human
  - Drawing the conclusions from data collected and analyzed
  - It is a difficult task and requires high degree of skills and experience
  - Correct interpretation will lead to a valid conclusion of study and this can aid one in taking suitable decisions

Functions of Statistics

- It provides facts in a definite form
- It simplifies mass of figures
- It facilitates comparisons
- It helps in formulating and testing hypothesis
- It helps in predictions
- It helps in formulating suitable policies

purpose

→ predictive analytics

↳ model

Limitations of Statistics

- Unless the data are properly collected and critically interpreted, there is every likelihood of drawing wrong conclusions
- Statistics does not deal with individual measurements
- Statistics deals only with Quantitative Characteristics
- Statistical results are true only on average
- Statistics is only one of the methods of studying a problem
- Statistics can be misused

Data Collection

50-60 % of project

Data Collection

- This is one of the most important and difficult tasks in Statistical Investigation
- Utmost care must be exercised while collecting the data as it constitute the foundation on which the superstructure of statistical analysis is built
- The results obtained from analysis are properly interpreted and policy decisions are taken
- Hence, if the data is inaccurate and inadequate, the whole analysis will be faulty and the decisions taken will be misleading
- Types
  - Primary Data — direct data — first hand data
    - Data collected by the researcher →
  - Secondary Data
    - Data collected by other agencies
      - → Kaggle

Primary Data

- Primary data are collected by a study, specifically designed to fulfill the data needs of the problem
- Such data are original in character and are generated in large number of surveys
- E.g. Census data collected by ministry of home affairs
- It is preferable to make use of primary data whenever possible because
  - The secondary source may contains mistakes due to the errors in transcriptions
  - Primary source frequently includes definitions of terms and units used
  - Primary source often includes schedule and description of procedures used in selecting data
  - Primary source often shows data in greater details
- Methods
  - Direct Personal Interviews
  - Indirect Oral Interviews
  - Information from correspondents
  - Mailed questionnaire method
  - Schedules sent through enumerators

→ digital surveys

→ scraping —

└ cron job

scraping

selenium / scrapy

data collection

requests

Direct Personal Interviews — most accurate results

- Conducting face-to-face interview with whom the information is to be obtained
- The interviewer asks the direct questions pertaining to the survey and collects desired answers
- Merits
  - Response is more encouraging as most people are willing to supply information when approached personally
  - It is more accurate as interviewer can clear the doubts of informants
  - It is also possible to collect supplementary information about informant's personal characteristics
  - Questions about which informant is likely to be sensitive can be carefully sandwiched between other questions
  - Language of the communication can be adjusted to the status and educational level of informants
- Demerits
  - It may be very costly where the number of people to be interviewed is large and they are spread over large area
  - The chances of personal prejudice and bias are greater
  - The interviewer must be thoroughly trained and supervised, otherwise they may not be able to obtain desired information
  - More time is required to collect information

Indirect Oral Interviews

- This approach is adopted in those cases where the information to be obtained is of a complex nature and informants are not inclined to respond if approached directly
- The Investigator contacts third parties called witnesses capable of supplying necessary information
- E.g. collecting information regarding addition to drugs, alcohols etc.
- Merits
  - This method is less costly than interviewing informants directly
- Demerits
  - The correctness of information obtained depends upon
    - Type of persons whose evidence is being recorded
    - Ability of interviewers to draw out the information from witness
    - Honesty of interviewers who are collecting the information
  - For the success of this method, it is necessary that the evidence of one person alone is not relied upon; views from number of persons should be considered to find out real position

Information from Correspondents

- In this methods, investigator appoints local agents or correspondents in different places to collect information
- These correspondents collect and transmit the data to central office where it is processed
- Newspapers generally adopt this method
- This method is also adopted by various government departments to collect regular information from wide area
- Merits
  - This method is very cheap and appropriate in extensive investigation
- Demerits
  - It may not produce accurate results because of personal prejudice and bias of correspodents

Mailed Questionnaire Method → (questions)

- Under this method, a list of questions pertaining to the survey (known as questionnaire) is prepared and sent to the various informants by post or by email
- The questionnaire contains list of questions and provides spaces for answers
- Request is made to the informants through a covering letter to fill up the questionnaire and send it back
- The questionnaire studies can be classified on the basis of
  - Degree to which the questionnaire is formalized or structured
  - Disguise or lack of disguise of the questionnaire
  - Communications method used
- Merits
  - Can be easily adopted where field of investigation is very vast and informants are spread over wide geographical area
  - It is relatively cheap
- Demerits
  - It can be adopted only where the informants are literate
  - It involves some uncertainty about responses
  - Information may not be correct and it may be difficult to verify accuracy

Secondary Data Sources → other agencies

- In most of the studies, investigator finds it impractical to collect first hand information on all related issues and as such use of the data collected by others
- There is a vast amount of published from which statistical studies may be made and fresh statistics are constantly in a a state of productions
- The sources of secondary data can be classified in
  - Published sources → websites, research papers, news papers, magazines
  - Unpublished sources
    - ↳ data acquisition

data collection

primary data

- direct personal interview
- oral interview
- from correspondents
- mail

secondary data

- published
- unpublished

Selection of appropriate method

- Nature, scope and object of enquiry
  - This constitutes the most important factor affecting the choice of a particular method
  - The method selected should be such that it suits the type of enquiry that is to be conducted by the researcher.
  - This factor is also important in deciding whether the data already available (secondary data) are to be used or the data not yet available (primary data) are to be collected
- Availability of funds
  - Availability of funds for research project determines to a large extent the method to be used for the collection of data
  - When funds at the disposal of the researcher are very limited, he will have to select a comparatively cheaper method which may not be as efficient and effective as some other costly method
  - Finance, in fact, is a big constraint in practice and the researcher has to act within this limitation
- Time factor
  - Availability of time has also to be taken into account in deciding a particular method of data collection
  - Some methods take relatively more time, whereas with others the data can be collected in a comparatively shorter duration
  - The time at the disposal of the researcher, thus, affects the selection of the method by which the data are to be collected

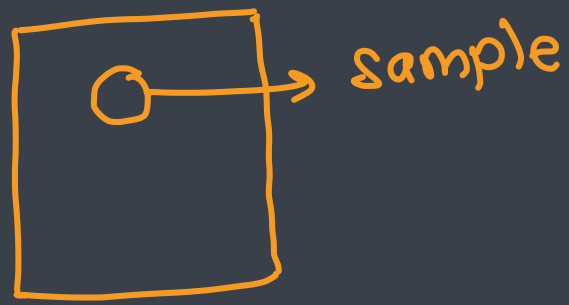
Data set
original

Population

Sampling

Sample

chosen few



Population (whole dataset)

- Generally, population refers to the people who live in a particular area at a specific time
- But in statistics, population refers to data on your study of interest
- A complete enumeration of all items in the 'population' is known as a census inquiry
- It can be presumed that in such an inquiry, when all items are covered, no element of chance is left and highest accuracy is obtained
- It can be a group of individuals, objects, events, organizations, etc.
- You use populations to draw conclusions
- An example of a population would be the entire student body at a school
  - It would contain all the students who study in that school at the time of data collection
  - Depending on the problem statement, data from each of these students is collected
  - An example is the students who speak Hindi among the students of a school

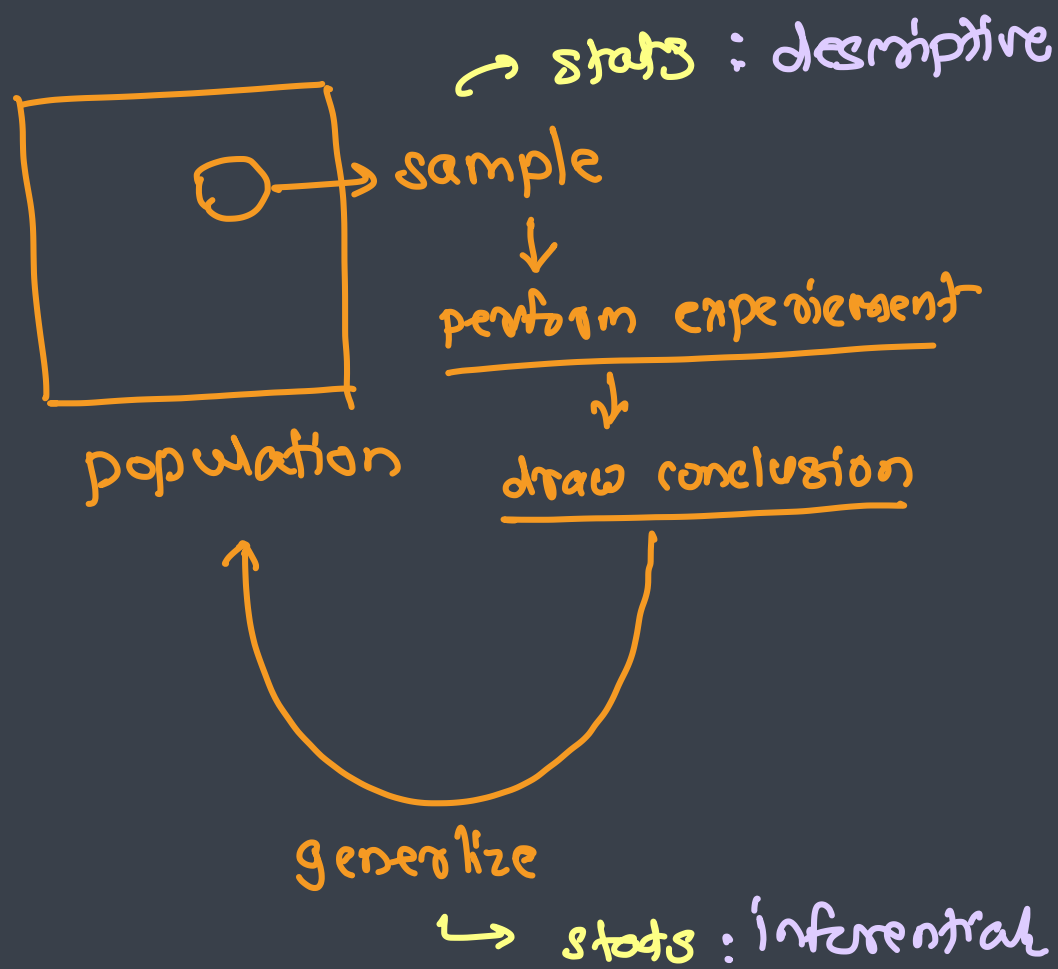
Census method

Every member of population is involved

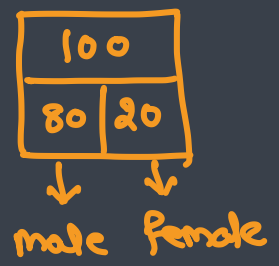
- **Merits**
  - Data are obtained from each and every unit of population
  - Results obtained are likely to be more representative, accurate and reliable
  - It is appropriate method of obtaining information on rare events
  - Data of complete enumeration census can be widely used as a basis for various surveys
- **Demerits**
  - Census method is not very popularly used in the practice
  - The effort, money and time required to carry out complete enumeration will generally be very large
  - In many cases it is not possible to contact every member of the census

Sampling

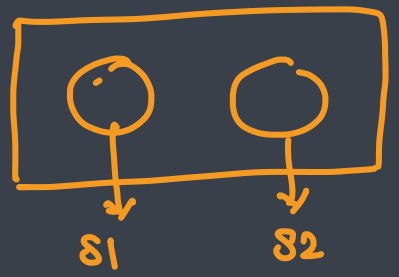
- It is the process of learning about the population on the basis of a sample drawn from it
- In this technique, instead of every unit in the population, only a part of it is studied and the conclusions are drawn on that basis for the whole population
- A sample is a subset of population
- The process of sampling involves three elements
  - Selecting the sample
  - Collecting the information from sample
  - Making the inference about the population



Essentials of Sampling

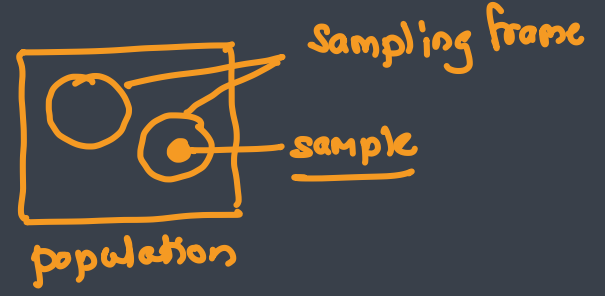


- **Representativeness** → there should not be any bias
  - A sample should be selected so that it truly represents the population
  - Otherwise the results obtained may be misleading
- **Adequacy**
  - The size of sample should be adequate, otherwise it may not represent the characteristics of the population
- **Independence**
  - All items of the sample should be selected independently of one another and all the items of the population should have the same chance of being selected in the sample → Random sampling
- **Homogeneity**
  - There is no basic difference in the nature of units of population and that of the sample
  - It two samples from the same population are taken, then they should give more or less same results



Steps in sample design

- **Type of universe (population)**
  - Clearly define the type of universe
  - It can be finite or infinite depending upon the type
- **Sampling unit**
  - A decision has to be taken concerning a sampling unit before selecting sample
  - Sampling unit may be a geographical one such as state, district, village, etc., or a construction unit such as house, flat, etc., or it may be a social unit such as family, club, school, etc., or it may be an individual
- **Source list**
  - It is also known as 'sampling frame' from which sample is to be drawn
  - It contains the names of all items of a universe
  - If source list is not available, researcher has to prepare it
  - Such a list should be comprehensive, correct, reliable and appropriate
  - It is extremely important for the source list to be as representative of the population as possible

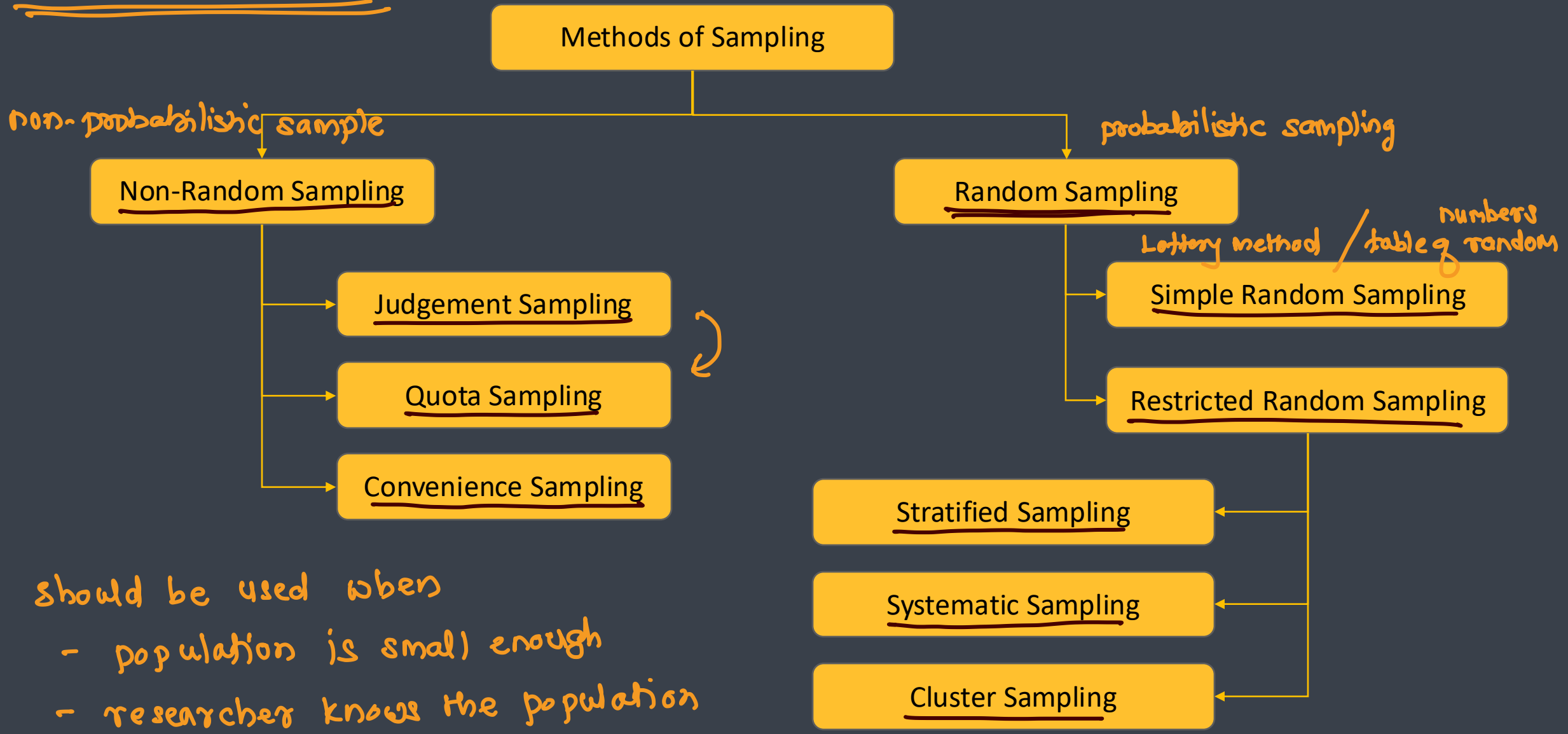


- **Budgetary constraint**
  - Cost considerations, from practical point of view, have a major impact upon decisions relating to not only the size of the sample but also to the type of sample

Steps in sample design

- **Size of sample**
  - This refers to the number of items to be selected from the universe to constitute a sample
  - This a major problem before a researcher
  - The size of sample should neither be excessively large, nor too small, it should be optimum
  - An optimum sample is one which fulfills the requirements of efficiency, representativeness, reliability and flexibility
- **Parameters of interest**
  - In determining the sample design, one must consider the question of the specific population parameters which are of interest
  - For instance, we may be interested in estimating the proportion of persons with some characteristic in the population, or we may be interested in knowing some average or the other measure concerning the population
- **Sampling procedure**
  - Researcher must decide about the technique to be used in selecting the items for the sample
  - In fact, this technique or procedure stands for the sample design itself
  - There are several sample designs out of which the researcher must choose one for his study

Methods of Sampling



should be used when
- population is small enough
- researcher knows the population very well known

Judgement Sampling

- The choice of sample items depends exclusively on the judgement of investigator
- Investigator exercises his judgment in the choice and includes those items in the sample which he thinks are the most typical of the population with required characteristics
- E.g. a sample of 10 students needs to selected from a class of 60, investigator would select 10 students who in his opinion are representative of the class
- Merits
  - If the population has very small number of units, judgement sampling may produce better result
  - When the investigator has good idea about the population, he can select the samples adequately
- Demerits
  - It is not scientific as the population units may be affected by the personal prejudice or investigator's bias
  - There is no objective way of evaluating reliability of sample results

Quota Sampling

- It is a type of Judgment sampling and is most commonly used in non-probability category
- Quotas are set up according to some specified characteristics like income category groups etc.
- Interviewer then told to interview the quota members
- Within the quota, the selecting of sample units depends on personal judgement
- It is often used in public opinion studies and it occasionally provides satisfactory results
- Merits
  - The cost per person interviewed may be relatively small for a quota sampling
- Demerits
  - The samples generated may have bias

Convenience Sampling

- A convenience sample is obtained by selecting convenient population units
- This process of sampling is also known as Chunk
- A Chunk refers to a portion of population being investigated
- E.g. a sample obtained from readily available list of registered cars or telephone directory etc.
- This method is generally used for making pilot studies
- Merits
  - It is very easy to select the sample using this process
- Demerits
  - The results obtained by following this method can hardly be representative of population
  - They are generally biased and unsatisfactory

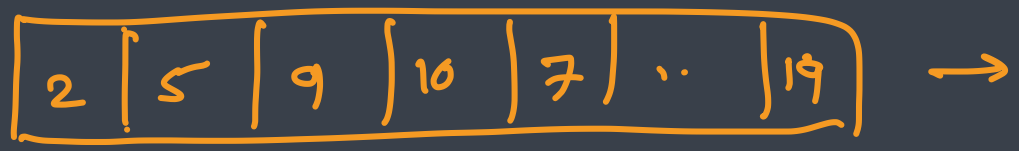
Probability Sampling Methods → Random Sampling

- **Merits**
  - It does not depends upon existence of detailed information about the population
  - It provides estimates which are essentially unbiased and have measurable precision
  - It is possible to evaluate the relative efficiency of various sample designs only when probability sampling is used
- **Demerits**
  - It requires high level of skill and experience for its use
  - It requires a lot of time to plan and execute probability sampling
  - The costs involved in probability sampling are generally larger as compared to non-probability sampling methods

→ Every member in the population has equal chance of being selected in the sample

Simple Random Sampling

- In this technique, each and every unit of population has equal opportunity of being selected
- Personal bias of investigator does not influence the selection
- "Random" does not mean haphazard or hit-or-miss, it rather means that the selection process is such that chance only determine which items will be included in the sample
- To ensure the randomness of selection there are two methods
  - Lottery method
    - This is very popular method of taking a random sample
    - While using the lottery method, it is absolutely essential to see that the slips are of identical size, shape and color, otherwise there is lot of possibility of personal prejudice and bias affecting the result
  - Table of random numbers
    - The lottery method becomes cumbersome if the size of population increases
    - In this method, the selection will be made based on the table which contains random numbers



Stratified Random Sampling → stratum based selection → characteristic of population

- It attempt to design better samples than simple random as it uses population information
- Procedure for stratified random sampling
  - Population is divided into groups which are mutually exclusive
  - A simple random sample is then chosen independently from each group

group = stratum

- This process differs from simple random in which random samples are selected from whole population
- Issues involved in selecting stratified random samples are
  - Base of stratification
  - Number of strata
  - Sample size within strata
- There are two ways to execute the stratified random sampling
  - → Proportional Stratified Random Sampling
  - → Disproportional Stratified Random Sampling

id	name	age	salary
1	x	20	50k
2	y	30	60k
3	z	21	70k
4	a	50	60k
1000	b	70	0

20-70

age

10%

Stratum / group

20-30	30-40	40-50	50-60	60-70
400	200	100	200	100
Random ⇒ 40	20	10	20	10

Strata

= 100 — sample

Stratified Random Sampling

- Merits
  - More representative
    - since the population is divided first into strata, the sample of each group represents the population better than simple random sampling
  - Greater accuracy
    - Accuracy is maximum if each stratum is so formed that it consists of uniform or homogeneous items
  - Greater geographical concentration
    - As compared to simple random, stratified sample can be more concentrated geographically
- Demerits
  - Utmost care must be exercised in dividing the population into various strata
  - The items from each stratus should selected at random which may be difficult in absence of skilled supervisor
  - Sample cost per observation may be quite high

Systematic Sampling → interval based selection

- A systematic sample if formed by selecting one unit at random and then selecting additional units at evenly spaced intervals until the sample has been formed
- This method is mostly used when the list of whole population is available
- The first item is selected randomly mostly using lottery method
- Subsequent items are selected by taking every kth item from the list where k refers to the sampling interval or sampling ration

1 (2) 3 .. 100

1, 10, 20, 30 ... 100

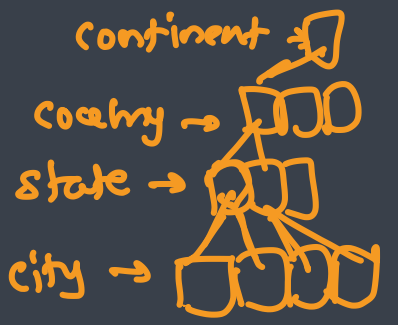
Sample

- Merits
  - It is simple and convenient to adopt
  - The time and work involved are relatively less
  - The results obtained are also found to be generally satisfactory provided cared is taken
- Demerits
  - It becomes less representative if the population has hidden periodicities
  - If the population is in a specific order then it is possible to select only specific items in the sample

Cluster Sampling

cluster → group → characteristic
↳ locality / area

- It is also known as Multi-Stage Sampling
- The random sampling is made of primary, intermediate and final units from a given population
- There are several stages in which the sampling process is carried out
- At first stage, units are sampled by some suitable method like simple random samples
- Then a sample second stage units is selected from each of the selected first stage units, again by some suitable method which may be the same or different from the method employed in the first stage
- Further stages may be added as per requirement
- **Merits**
  - Introduces flexibility in sampling method which is lacking in other methods
  - As one stage is completed, the method deals with less item which speeds up the process
- **Demerits**
  - In general, it is less accurate than a sample containing the same number of final stage units which have been selected by some single stage process



Sample Size → adequate size of sample should be selected for reliable results

- Size of sample is nothing but no of items in the selected sample
- Following factors should be considered while deciding sample size
  - **Population size**: Larger the size of population, bigger should be the size of sample
  - **Resources available**: If large resources are available, sample with bigger size should be selected
  - **Degree of accuracy or precision desired**: Greater the degree of accuracy is desired, larger should be the sample size
  - **Homogeneity or heterogeneity of population**: If population consists of homogeneous units, small sample may serve the purpose, but if the population consists of heterogeneous units, larger sample size should be selected
  - **Nature of Study**: For an intensive and repeated study, small sample may be suitable, but if the studies are not likely to be repeated, larger sample size should be preferred
  - **Method of sampling**: The size of sample is also influenced by method of sampling
  - **Nature of respondents**: Where it is expected a large number of respondents will not cooperate and send back the questionnaire, a large sample should be selected

Sample Size

$$n = \left(\frac{Z\sigma}{d}\right)^2$$

- Where
  - n = Sample Size
  - Z = Value at a specified level of confidence or desired degree of precision
  - σ = Standard deviation of population
  - d = Difference between Population mean and sample mean

- Example: Determine the sample size if standard deviation of population is 6, population mean = 25, sample mean = 23 and desired degree of accuracy is 99% (consider Z = 2.576)

$\sigma = 6$, d = population mean − sample mean = 25 − 23 = 2

$Z = 2.576$ $\quad n = \left(\frac{Z\sigma}{d}\right)^2 = \left(\frac{2.576 \times 6}{2}\right)^2 = 59.1 \approx 60$

Sampling

- Merits
  - Less Time consuming
  - Less Costly
  - More reliable results
  - More detailed information
  - Sampling method is the only method that can be used in certain cases
  - Sample method is often used to judge the accuracy of information obtained on population
- Demerits
  - A sample survey method must be carefully planned and executed, otherwise the results may be inaccurate
  - Sampling generally requires expert services
  - At times, sampling plan may be so complicated that it requires more time, labor and money than completed count
  - If information about each and every unit is available then sampling is not required

Sampling Errors

- Even if utmost care has been taken in selecting a sample, the results derived from a sample study may not be exactly equal to the true value in the population
- The reason is that, the estimation is based on the part (sample) and not on the whole
- Hence the sample gives rise to the sample errors, also known as sampling fluctuations
- Types
  - Biased Errors
    - These errors arise from bias in the selection, estimation etc.
    - Causes
      - Faulty Process of selection
      - Faulty work during the collection
      - Faulty methods of analysis
  - Unbiased Errors
    - These errors arise due to the chance difference between the members of population included in sample and those which are not included


Data

① Set the purpose of exp. → hypothesis

② know / indentify the population

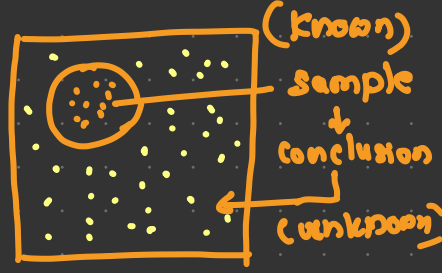
③ select sample using required method ↷ analysis

④ perform the experient [describe the data]

⑤ draw the conclusion [based on sample]

⑥ generalize the conclusion [apply on whole population]

[a test which will decide whether the conclusion can be generalized]



Statistics types

Descriptive Statistics	Inferential Statistics
Concerned with describing the target population	Make inferences from sample and generalize them for the entire population
Organize, analyze and present the data in meaningful manner → more understanding of data	Compare, test and predict the future outcomes → p-value
Final results are shown in the form of charts, tables or graphs → visualization	Final result is probability score → whether the conclusion can be generalized
Describes the data already known → sample post data	Tries to make the conclusion about the population that is beyond the data available
Tools: - Measures of central tendency - Measures of dispersion	Tools - Hypothesis testing - Analysis of variance

Individuals and Variables

- Meaning outside the statistics
- Individuals
  - Are the people
  - E.g. 60 students for the course
- Variable
  - Is the factor that can vary
  - E.g. time the shop can make cake

data

- Meaning inside the statistics
- Individuals → Record / Row
  - Objects included in the study
  - E.g. records, people, objects
- Variable
  - Characteristics of individuals to be measured or observed
  - E.g. age or name of the person

Variable
- Random Variable
- column
- attribute
- feature

Individuals and Variables

variable, features

individual

Row

Name	Email	Phone	Address	Age
Person1	p1@test.com	7845343456	Pune	56
Person2	p2@test.com	6644664466	Mumbai	67
Person3	p3@test.com	7676767676	Satara	30
Person4	p4@test.com	8989898989	Karad	50

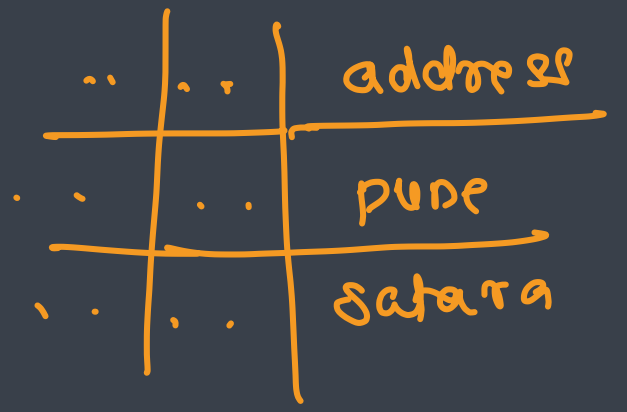
Qualitative

Quantitative

Variables: Qualitative → textual

- Variables that are not measurement variables
- Also known as categorical variables *
- Take category or label values and place an individual into one of several groups
- Each observation can be placed in only one category — mutually exclusive
- The categories are mutually exclusive
- E.g. political party, profession, gender, whether person smokes
- Types
  - Nominal
    - Can not be ordered from smallest to largest
  - Ordinal
    - Can be arranged in order of categories
    - But difference between data values can no be calculated or meaningless

pune, satara



feedback : sad unhappy happy awesome

Variables: Quantitative → Numeric

- Variables whose values result from counting or measuring something
- Also known as continuous variable
- Take numerical values and represent some kind of measurement
- E.g. height, weight, age
- Types
  - Interval
    - There is no true zero
  - Ratio
    - There is a true zero

Qualitative data is NOT supported in stats

→ Qualitative → Quantitative

one hot encoding

label encoding

Quantitative

temperature : 20, 21, 20, 28, 29, 20

Qualitative / categorical → colors : red, purple, red, green

Quantitative / categorical → result : 0, 1, 0, 1, 0, 0, 0, 1, 1

categories = [0, 1]

Series

value

marks = 20, 21, 24 . . .
series

- Collection of data points is called as a series
- Types
  - Time Series
    - A series of data that is arranged chronologically, or in relation to time

    → second, min, hr, . . . → stock prediction
  - Frequency Series
    - A series of data that is formed along with the frequencies of their occurrences
    - Types
      - Individual series
      - Discrete series
      - Continuous series

Individual Series

- Each value of the variable occurs for only once
- The frequency of occurrence of all the values in such a series is only one
- Such series are displayed without the frequency column
- E.g.
  - Marks: 40 60 50 80 45 85 90 67
  - Ages: 30 40 35 45 56 60

value	frequency
30	1
40	1
35	1
45	1
56	1
60	1

Discrete series

- The different values of a variable are shown in a discontinuous manner along with their respective frequencies
- Such a series can also be arranged either in ascending, or in descending order
- E.g.

frequency table

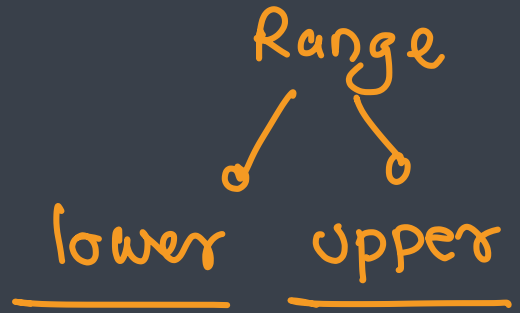
Marks	# Students
80	2
40	4
45	3
60	6
86	3
90	3
70	10

Wages	# Workers
150	10
160	20
170	10
200	6
500	4
1000	2
1500	1

80, 80, 40, 40, 40, 40, 45, 45, 45 .....

Continuous Series

- Different values of the variables are stated in a continuous manner along with their respective frequencies
- Such series can be stated either in the form exclusive, or in the from of inclusive class intervals along with their respective class frequencies
- E.g.



Marks Range	# Students
10-20	5
20-30	0
30-40	20
40-50	15
50-60	10
60-70	15
70-80	10
80-90	20
90-100	5

class