EMBA Program
MB-511

Data Science for Managerial Decisions (MB 511)

- Building Data Science Solutions
- Data Collection and Data Pre-Processing

Instructor
Anant Prakash Awasthi

Building a Data Science Solution

Approaches – Methodologies

Methodologies play a crucial role in building data science solutions by providing structure, guiding decision-making, and ensuring that projects are executed efficiently and effectively.

- Providing a Structured Framework
- Facilitating Collaboration
- Ensuring Alignment with Business Objectives
- Enhancing Data Quality and Preparation
- Improving Model Performance and Evaluation
- Managing Complexity and Risk
- Fostering Iteration and Continuous Improvement
- Guiding Deployment and Operationalization
- Ensuring Accountability and Documentation
- Optimizing Resource Allocation
- Improving Stakeholder Engagement and Communication
- Supporting Scalability and Reproducibility

Data Science Project Execution Methodologies

- **CRISP-DM (Cross-Industry Standard Process for Data Mining)**
- SEMMA (Sample, Explore, Modify, Model, Assess)
- OODA Loop (Observe, Orient, Decide, Act)
- KDD (Knowledge Discovery in Databases)
- Team Data Science Process (TDSP)
- Lean AI/Agile Data Science
- Google's TFX (TensorFlow Extended) Pipeline for Machine Learning
- ASUM-DM (Analytics Solutions Unified Method for Data Mining)
- Design Thinking for Data Science
- End-to-End Machine Learning Pipeline

CRISP-DM (Cross-Industry Standard Process for Data Mining)

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely used, industry-standard methodology for data science projects. Developed in the late 1990s, it provides a structured approach to building data-driven solutions, particularly in the context of data mining and predictive modeling. CRISP-DM is highly adaptable and is applied across a variety of industries, from finance to healthcare, retail, and more.

Why CRISP-DM:

- Industry-Agnostic
- Iterative and Non-Linear Process
- Emphasizes Business Understanding
- Data-Centric
- Model Agnostic
- Evaluation-Focused
- Deployment and Monitoring
- Well-Established and Widely Used
- Phases Are Well-Defined and Modular
- Supports Both Supervised and Unsupervised Learning
- Encourages Documentation and Transparency
- Focuses on Practical Implementation

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Characteristics

Industry-Agnostic

- CRISP-DM is designed to be adaptable to any industry, from finance to healthcare, retail, and beyond. It provides a general framework that can be applied to various types of data mining and machine learning problems without being tied to a specific domain.
- Benefit: It can be used in diverse fields and is highly versatile, allowing for broad application.

Iterative and Non-Linear Process

- Although CRISP-DM outlines a structured sequence of six phases, the process is iterative rather than strictly linear. Data scientists often need to loop back to previous stages based on new insights or issues that arise during later stages.
- Benefit: Encourages continuous improvement and refinement of models and approaches, ensuring that the solution evolves with the project.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Characteristics

Emphasizes Business Understanding

- CRISP-DM places significant importance on understanding the business problem before diving into data. The very first phase, "Business Understanding," ensures that the project aligns with real-world business objectives and that the solution delivers practical value.
- Benefit: Keeps the focus on solving business problems, ensuring the project’s outcome is useful and relevant to stakeholders.

Data-Centric

- The methodology emphasizes the importance of data at every step. It has dedicated phases for understanding, preparing, and exploring data, which ensures that data is treated as a crucial asset.
- Benefit: Forces data scientists to thoroughly explore and prepare data, leading to higher-quality models and insights.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Characteristics

Model Agnostic

- CRISP-DM does not prescribe specific modeling techniques or algorithms. It allows for the use of any method that fits the problem at hand, whether it is regression, decision trees, clustering, or deep learning.
- Benefit: Provides flexibility in choosing the best modeling approach for the given problem.

Evaluation-Focused

- Before moving forward to deployment, CRISP-DM emphasizes rigorous evaluation of the model's performance against the business objectives. This ensures that the solution not only works from a technical perspective but also meets the predefined success criteria.
- Benefit: Ensures that only well-performing models are deployed, reducing the risk of failure in production environments.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Characteristics

Deployment and Monitoring

The final phase of CRISP-DM includes deploying the model into production and monitoring its performance over time. This phase ensures that the solution is fully operational and integrated into the organization’s workflows.

Benefit: Helps transition from a prototype model to a fully functional, real-world solution that delivers continuous value.

Well-Established and Widely Used

CRISP-DM is one of the most established and commonly used methodologies in the field of data science. Its widespread adoption means that many professionals are familiar with it, making it easier for teams to collaborate and follow a shared process.

Benefit: Reduced learning curve and increased compatibility across teams and projects.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Characteristics

Phases Are Well-Defined and Modular

Each of the six phases in CRISP-DM—Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment—is clearly defined, which makes it easy to know what tasks to focus on at any given point.

Benefit: Provides clarity and structure, making it easy to plan and manage projects effectively.

Supports Both Supervised and Unsupervised Learning

CRISP-DM can be applied to a wide range of data science tasks, whether the goal is to predict outcomes (supervised learning) or find patterns in data (unsupervised learning).

Benefit: Versatile for various types of projects, from predictive modeling to clustering and segmentation.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Characteristics

Encourages Documentation and Transparency

CRISP-DM encourages thorough documentation at each stage, ensuring that the project is well-documented and that key decisions are recorded. This helps ensure transparency and traceability, especially when models need to be retrained or audited.

Benefit: Facilitates knowledge sharing, reproducibility, and accountability within data science teams.

Focuses on Practical Implementation

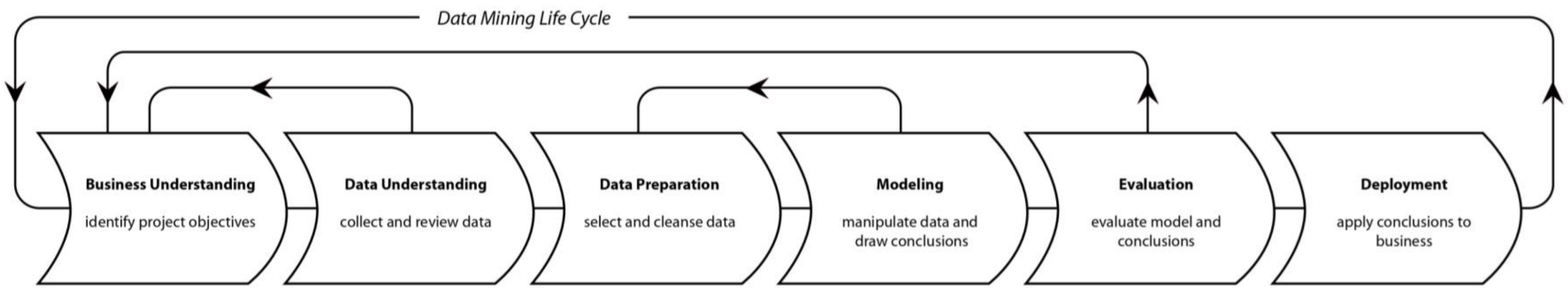
CRISP-DM is not just about theoretical models or academic research. The final phase—deployment—ensures that models are implemented and deliver tangible results for the business.

Benefit: Encourages real-world application and focuses on delivering value rather than just building models.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Workflow

EMBA Program
MB-511



Determine Business Objectives
Background
Business Objectives
Business Success Criteria
(Log and Report Process)

Assess Situation
Inventory of Resources, Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
(Log and Report Process)

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
(Log and Report Process)

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques
(Log and Report Process)

Collect Initial Data
Initial Data Collection Report
(Log and Report Process)

Describe Data
Data Description Report
(Log and Report Process)

Explore Data
Data Exploration Report
(Log and Report Process)

Verify Data Quality
Data Quality Report
(Log and Report Process)

Data Set
Data Set Description
(Log and Report Process)

Select Data
Rationale for Inclusion/ Exclusion
(Log and Report Process)

Clean Data
Data Cleaning Report
(Log and Report Process)

Construct Data
Derived Attributes
Generated Records
(Log and Report Process)

Integrate Data
Merged Data
(Log and Report Process)

Format Data
Reformatted Data
(Log and Report Process)

Select Modeling Technique
Modeling Technique
Modeling Assumptions
(Log and Report Process)

Generate Test Design
Test Design
(Log and Report Process)

Build Model Parameter Settings
Models
Model Description
(Log and Report Process)

Assess Model
Model Assessment
Revised Parameter
(Log and Report Process)

Evaluate Results
Align Assessment of Data Mining Results with Business Success Criteria
(Log and Report Process)

Approved Models
Review Process
Review of Process
(Log and Report Process)

Determine Next Steps
List of Possible Actions
Decision
(Log and Report Process)

Plan Deployment
Deployment Plan
(Log and Report Process)

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
(Log and Report Process)

Produce Final Report
Final Report
Final Presentation
(Log and Report Process)

Review Project
Experience
Documentation
(Log and Report Process)

Generic Tasks
Specialized Tasks
(Process Instances)

a visual guide to CRISP-DM methodology

SOURCE CRISP-DM 1.0
http://www.crisp-dm.org/download.htm
DESIGN Nicole Leaper
http://www.nicoleleaper.com

Data Collection and Data Pre-Processing

Data Science for Managerial Decisions (MB 511)

EMBA Program
MB-511

Program Overview

- Introduction to Data Science
- Information Technology An Overview
- Applications of Data Science in various fields
- MIS and Control Systems
- Data Collection and Data Pre-Processing
- Building Information Systems
- Support Systems for Management Decisions

Data Collection and Data Pre-Processing

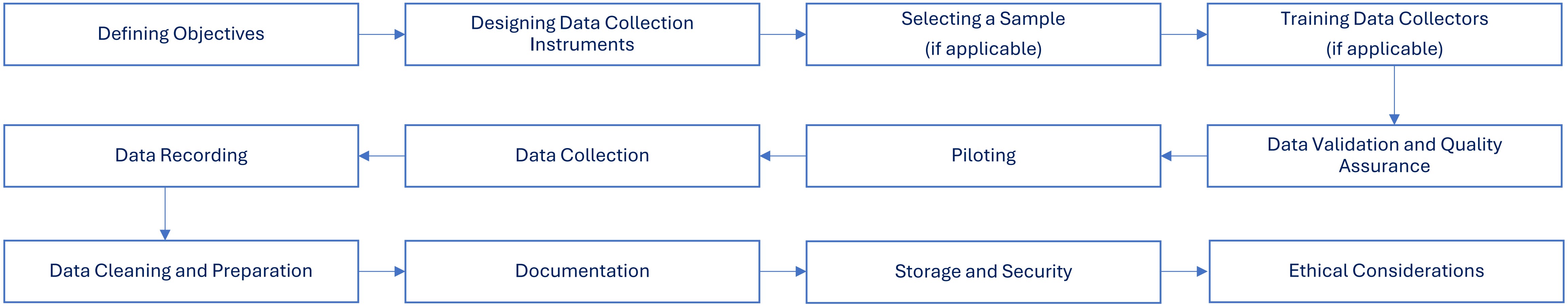
- Introduction to Data Collection
- Methods of Data Collection in Management
- Designing Data Collection Instruments
- Sampling Techniques
- Data Collection Planning and Management
- Understanding Data Pre-Processing
- Data Cleaning Techniques
- Quality Assurance in Data Pre-Processing

Data Collection and Data Pre-Processing

Introduction to Data Collection

Data collection is the process of gathering, measuring, and recording information systematically. It is a fundamental step in the research process, whether for scientific research, market research, social science, or any other field where data plays a crucial role in decision-making, analysis, and understanding.

The Process



Defining Objectives
Designing Data Collection Instruments
Selecting a Sample (if applicable)
Training Data Collectors (if applicable)
Data Validation and Quality Assurance
Piloting
Data Collection
Data Recording
Data Cleaning and Preparation
Documentation
Storage and Security
Ethical Considerations

Data Collection and Data Pre-Processing

Introduction to Data Collection

1. Defining Objectives: Clearly defining the purpose of data collection is essential. Understanding what information is needed and why it's needed helps guide the entire process.
2. Designing Data Collection Instruments: This step involves determining what methods and tools will be used to collect data. Common data collection instruments include surveys, interviews, questionnaires, observation forms, and experiments.
3. Selecting a Sample (if applicable): If the data collection involves sampling, determining the appropriate sampling method (random sampling, stratified sampling, etc.) and selecting the sample is crucial. The sample should be representative of the population being studied.

Data Collection and Data Pre-Processing

Introduction to Data Collection

4. Training Data Collectors (if applicable): If multiple individuals will be involved in data collection, ensuring that they are adequately trained on the data collection instruments and procedures is essential to maintain consistency and reliability.

5. Piloting: Before conducting full-scale data collection, it's often beneficial to pilot the data collection instruments and procedures to identify any potential issues and make necessary adjustments.

6. Data Collection: This is the actual process of gathering data according to the defined procedures. Data collection can involve various methods such as surveys, interviews, observations, experiments, and data extraction from existing sources (e.g., databases, records).

Data Collection and Data Pre-Processing

Introduction to Data Collection

7. Data Recording: Once data is collected, it needs to be recorded in a systematic and organized manner. This may involve entering data into spreadsheets, databases, or other data management systems.

8. Data Validation and Quality Assurance: After data collection, it's important to validate the data to ensure accuracy and reliability. This may involve checking for errors, inconsistencies, or missing values and implementing quality control measures.

9. Data Cleaning and Preparation: Data collected may require cleaning and preparation before analysis. This involves tasks such as handling missing data, coding variables, transforming data, and checking for outliers.

Data Collection and Data Pre-Processing

Introduction to Data Collection

10. Documentation: Proper documentation of the data collection process, including details about methods, procedures, and any issues encountered, is essential for transparency, reproducibility, and future reference.
11. Storage and Security: Ensuring that collected data is stored securely and confidentially is crucial to protect the privacy and integrity of the data.
12. Ethical Considerations: Data collection should be conducted in accordance with ethical principles and guidelines, ensuring that participants' rights and confidentiality are respected.

Data Collection and Data Pre-Processing

Introduction to Data Collection

Data collection and processing laws vary significantly from country to country and are influenced by factors such as cultural norms, historical context, legal traditions, and government policies.

1. European Union (EU) - General Data Protection Regulation (GDPR)
2. United States - Health Insurance Portability and Accountability Act (HIPAA)
3. Canada - Personal Information Protection and Electronic Documents Act (PIPEDA)
4. Brazil - Lei Geral de Proteção de Dados (LGPD)
5. India - The Personal Data Protection Bill (PDPB)
6. China - Cybersecurity Law
7. Australia - Privacy Act 1988

Data Collection and Data Pre-Processing

Methods of Data Collection in Management

In management, data collection methods play a vital role in gathering information to support decision-making, problem-solving, and strategic planning. There are several data collection methods commonly used in management.

1. Surveys
2. Interviews
3. Observation
4. Document Analysis
5. Focus Groups
6. Case Studies
7. Action Research
8. Experiments
9. Metrics and Key Performance Indicators (KPIs)

Data Collection and Data Pre-Processing

Methods of Data Collection in Management

- Surveys: Surveys involve asking individuals or groups a series of questions to gather information about their opinions, preferences, behaviors, or experiences. Surveys can be conducted through various means, including online surveys, paper surveys, telephone interviews etc.
- Interviews: Interviews involve direct interaction between the interviewer and the interviewee to gather detailed information. Interviews can be structured (with predetermined questions) or unstructured (more open-ended), depending on the research objectives. They can be conducted one-on-one or in group settings.
- Observation: Observation involves systematically watching and recording behaviors, activities, or events in real-time. This method can provide valuable insights into organizational processes, interactions, and work environments. Observations can be conducted covertly (without participants' knowledge) or overtly (with their knowledge).

Data Collection and Data Pre-Processing

Methods of Data Collection in Management

- Document Analysis: Document analysis involves examining existing documents, records, reports, or other written materials to extract relevant information. This method can include reviewing financial statements, meeting minutes, policies, procedures, emails, and other written communications.
- Focus Groups: Focus groups involve bringing together a small group of individuals to participate in a facilitated discussion about a specific topic or issue. Focus groups encourage interaction and idea generation among participants and can provide in-depth insights into attitudes, perceptions, and preferences.
- Case Studies: Case studies involve in-depth analysis of a particular organization, project, event, or situation over time. Case studies often combine multiple data collection methods, such as interviews, document analysis, and observation, to provide a comprehensive understanding of the subject.

Data Collection and Data Pre-Processing

Methods of Data Collection in Management

- Action Research: Action research involves conducting research in collaboration with practitioners to address specific organizational challenges or problems. This method emphasizes participation, reflection, and iterative problem-solving cycles to generate actionable insights and drive organizational change.
- Experiments: Experiments involve manipulating variables in a controlled setting to observe their effects on outcomes. While less common in management research compared to other fields, experiments can be used to test hypotheses, evaluate interventions, or measure causal relationships.
- Metrics and Key Performance Indicators (KPIs): Metrics and KPIs involve collecting quantitative data on organizational performance indicators such as sales, revenue, customer satisfaction, employee productivity, and operational efficiency. This data is often collected continuously or periodically through organizational systems and processes.

Data Collection and Data Pre-Processing

Designing Data Collection Instruments

Designing data collection instruments is a critical step in the research process, as it directly impacts the quality and reliability of the data collected. A well-designed instrument ensures that the data collected are accurate, relevant, and aligned with the research objectives.

1. Define Research Objectives
2. Select Data Collection Method
3. Develop Clear and Specific Questions
4. Consider Response Options
5. Pilot Test the Instrument
6. Consider the Context and Setting
7. Ensure Validity and Reliability
8. Ethical Considerations

Data Collection and Data Pre-Processing

Designing Data Collection Instruments

- Define Research Objectives: The first step in designing data collection instruments is to clearly define the research objectives and questions. Understanding what information needs to be collected and why it's needed helps guide the design process.
- Select Data Collection Method: Based on the research objectives, select the most appropriate data collection method(s) such as surveys, interviews, questionnaires, observations, experiments, or existing data sources. Consider factors such as the nature of the data, the target population, and resource constraints.
- Develop Clear and Specific Questions: For surveys, interviews, and questionnaires, develop clear, specific, and unambiguous questions that address the research objectives. Use simple language and avoid jargon or technical terms that may confuse respondents. Ensure that questions are relevant and directly related to the research objectives.

Data Collection and Data Pre-Processing

Designing Data Collection Instruments

- Consider Response Options: For closed-ended questions, provide response options that cover all possible answers and are mutually exclusive. Use appropriate scales (e.g., Likert scales, multiple-choice, yes/no) based on the nature of the data and the research objectives. For open-ended questions, allow respondents to provide detailed and qualitative responses.
- Pilot Test the Instrument: Before using the instrument for data collection, pilot test it with a small sample of participants to identify any issues or problems. This allows for refining and improving the instrument before full-scale implementation. Pay attention to the clarity of the questions, the length of the instrument, and the ease of completion.
- Consider the Context and Setting: When designing data collection instruments, consider the context and setting in which data will be collected. Adapt the instrument to the cultural, linguistic, and environmental factors relevant to the target population. Ensure that the instrument is appropriate for the intended respondents and the data collection environment.

Data Collection and Data Pre-Processing

Designing Data Collection Instruments

- Ensure Validity and Reliability: Ensure that the data collection instrument is valid (measures what it intends to measure) and reliable (produces consistent results). Use established techniques such as content validity, criterion validity, construct validity, and test-retest reliability to assess and enhance the quality of the instrument.

- Ethical Considerations: Consider ethical considerations such as informed consent, privacy, confidentiality, and anonymity when designing data collection instruments. Ensure that respondents are fully informed about the purpose of the research, their rights, and how their data will be used and protected.

Data Collection and Data Pre-Processing

Sampling Techniques

Sampling is the process of selecting a subset of individuals or items from a larger population for the purpose of collecting data and making inferences about the population as a whole. In other words, instead of collecting data from every member of the population, researchers select a sample that represents the population and use the data collected from the sample to draw conclusions about the entire population.

A good sample is one that accurately represents the characteristics of the population from which it is drawn. It should be selected using a valid sampling method, have a sufficient sample size, and minimize sampling bias. A good sample allows for reliable statistical inference and generalization to the larger population.

Data Collection and Data Pre-Processing

Sampling Techniques

In sample surveys, various sampling techniques are employed to select a subset of individuals or units from a larger population for data collection. Each sampling technique has its own advantages, limitations, and suitability for different research objectives and populations.

Simple Random Sampling (SRS):

- In simple random sampling, each individual or unit in the population has an equal chance of being selected for the sample.
- This method is straightforward to implement and ensures that each member of the population has an equal probability of inclusion.
- It requires a complete list of the population (sampling frame) and is suitable when the population is relatively homogeneous.

Data Collection and Data Pre-Processing

Sampling Techniques

Stratified Sampling:

- In stratified sampling, the population is divided into homogeneous subgroups called strata, and samples are independently selected from each stratum.
- This method ensures that each subgroup is adequately represented in the sample, leading to more precise estimates, especially when there is variability within the population.
- It requires knowledge of the population characteristics to create meaningful strata.

Systematic Sampling:

- Systematic sampling involves selecting every nth individual from a list of the population after randomly selecting a starting point.
- It is easier to implement than simple random sampling and can be more efficient when there is a natural ordering in the population.
- However, it may introduce bias if there is periodicity or clustering in the population.

Data Collection and Data Pre-Processing

Sampling Techniques

Cluster Sampling:

- In cluster sampling, the population is divided into clusters, and a random sample of clusters is selected. Then, all individuals within the selected clusters are included in the sample.
- This method is useful when it is impractical or costly to sample individuals directly, and when clusters naturally occur in the population.
- It can lead to increased sampling variability compared to other methods, especially if clusters are heterogeneous.

Multistage Sampling:

- Multistage sampling combines two or more sampling methods in successive stages. For example, clusters may be sampled using cluster sampling, and then individuals within selected clusters are sampled using simple random sampling or another method.
- This method is often used for large-scale surveys and allows for efficient sampling of large and diverse populations.
- It requires careful planning and coordination of multiple sampling stages.

Data Collection and Data Pre-Processing

Sampling Techniques

Probability Proportional to Size (PPS) Sampling:

- In PPS sampling, the probability of selecting a unit is proportional to its size or measure of importance in the population.
- This method is useful when the population units vary widely in size, and it ensures that larger units have a higher chance of being selected.
- It requires accurate information on the size of population units, which may not always be available.

Convenience Sampling:

- Convenience sampling involves selecting individuals who are readily available or easily accessible to the researcher.
- It is quick, inexpensive, and convenient but may lead to biased results if the sample does not accurately represent the population.
- It is often used in exploratory research or situations where other sampling methods are impractical.

Data Collection and Data Pre-Processing

Sampling Techniques

Snowball Sampling:

- Snowball sampling starts with an initial set of individuals who meet the inclusion criteria, and then additional participants are recruited through referrals from existing participants.
- It is useful for sampling hard-to-reach or hidden populations and can lead to the identification of rare or specialized groups.
- It may result in biased samples if referrals are not representative of the population.

Each sampling technique has its own strengths and weaknesses, and the choice of method depends on factors such as the research objectives, the nature of the population, resource constraints, and the desired level of precision and representativeness. In practice, researchers often use a combination of sampling techniques or employ specialized techniques tailored to specific research contexts.

EMBA Program
MB-511

Have a question?

Feel Free to Reach out at

- +91-88846-92929 (WhatsApp)
- anant.awasthi@outlook.com (E-Mail)