Content from different reference book suggested by Dr. Anita Morey Mam and mentioned in CDA Syllabus file, Presentations, Notes provided in 2023

NOTES ON BASIS OF CDA PPT SAMPLING AND SIMULATION 51 SLIDES

Data Sampling Notes:

- Data Sampling Basics:
  - Data sampling is a statistical technique used to select, manipulate, and analyze a subset of data points from a larger dataset to identify patterns and trends.
  - It enables data scientists, predictive models, and analysts to work with a manageable amount of data while still producing accurate findings.

- Advantages and Challenges:
  - Sampling is useful for large datasets that are impractical to analyze entirely, such as in big data analytics or surveys.
  - It's more efficient and cost-effective to analyze a representative sample than the entire dataset.
  - Size of the sample is important; sampling error can occur if the sample size is too small.
  - Sometimes, a small sample reveals critical information, while a larger sample might better represent the overall data but could be harder to manage.

- Types of Sampling Methods:
  - Probability Sampling:
    - Simple Random Sampling: Randomly selecting subjects from the entire population.
    - Stratified Sampling: Creating subsets based on a common factor and randomly sampling from each subgroup.
    - Cluster Sampling: Dividing the dataset into clusters based on a factor, then randomly sampling clusters for analysis.
    - Multistage Sampling: Similar to cluster sampling, involving multiple levels of clustering and sampling.
    - Systematic Sampling: Selecting samples at a regular interval from the population.
  - Nonprobability Sampling:
    - Sampling is determined by analyst judgment, making it harder to ensure representativeness.

- Sampling in Data Science:
  - In most studies, analyzing the entire population is challenging, so researchers use samples.
  - Different sampling methods introduce biases; understanding implications is crucial.
  - Two main categories: probability and non-probability sampling.
  - Probability sampling ensures each element has a known, non-zero chance of being in the sample.
  - Non-probability sampling might not represent the population well, but it can be cheaper or more feasible.

- Probability Sampling Methods:
  - Simple Random Sampling without Replacement (SRSWR): Randomly selecting elements until the desired sample size is reached.
  - SRSWR is unbiased, but a purely random sample might not always be representative.

  - Poisson Sampling: Elements go through Bernoulli trials to determine inclusion in the sample.
  - Bernoulli sampling is a special case when probabilities are the same for all elements.
  - Can result in random-sized samples.
  - Requires a list of all population elements.

These notes cover the fundamentals of data sampling, its advantages, challenges, and different methods, including probability and non-probability sampling approaches. It's important to understand the implications of different sampling designs to ensure accurate and meaningful analysis.

Data Sampling and Simulation Notes:

- Stratified Sampling:
  - Useful when population needs to be divided based on certain features.
  - Helps ensure representation of various groups within the sample.
  - Example: Surveying company employees for job satisfaction, stratifying by department to avoid bias.

- Benefits of Stratified Sampling:
  - Works well when variability within strata is small and variability between strata is significant.
  - Enhances accuracy by accounting for differences in different segments of the population.

- Challenges and Implementation:
  - Can be expensive and complex due to the need for prior information about the population.
  - Useful for intermediate studies between broader ones, utilizing existing data to guide smaller studies.

- Non-probability Sampling:
  - Volunteer Sampling: Gathering data from individuals who choose to participate, leading to potential bias.
  - Judgement Sampling: Selecting participants based on existing domain knowledge, prone to biases.

- Understanding Sampling Designs:
  - Crucial for data scientists to grasp different sampling designs and their implications.
  - Survey sampling is a specialized field, essential for statisticians and researchers.

- Simulation Overview:
  - Data simulation involves mirroring real-world conditions to predict, guide decisions, or validate models.
  - Different forms for different purposes: approximating known conditions, experimenting with scenarios, climate projections, etc.

- Simulation Features:
  - Graphical user interface for accessibility and ease of use.
  - Model building supported by adequate compute power and scalability.
  - Analytics integration and data import/export functionalities.

- Simulation Benefits and Uses:
  - Models behavior across complex systems.

  - Provides realistic models for prediction and validation.
  - Visualizes trends, aids decision-making, and guides strategy.
  - Used in industries like oil and gas, climate projections, and digital twin development.

- Data Simulation Software:
  - Various simulation tools available, tailored to different industries and purposes.

- Modelling and Simulations in Data Science:
  - Addressing the limitation of constant need for new data in machine learning.
  - Simulation models: mathematical and process models.
  - Used in various fields, including finance, medical training, epidemiology, and predictive analytics.

- Simulation and Predictive Analytics:
  - Both require models but serve different purposes.
  - Decision trees vs. machine learning: choice depends on system complexity and data availability.

These notes provide insights into data sampling methods, the benefits of stratified sampling, non-probability sampling approaches, the concept and applications of data simulation, and the role of simulation in predictive analytics.

Observational Sampling Design Notes:

- Observational vs. Experimental Studies:
  - Observational studies collect data by observing events, while experiments involve researchers assigning variables.
  - Causal conclusions are reasonable in experiments but risky in observational studies; they generally show associations.

- Causation and Observational Data:
  - Causal conclusions based on observational data can be misleading due to confounding variables.
  - Confounding variables are correlated with both explanatory and response variables, introducing bias.
  - Exhaustively searching for confounding variables is challenging and may not cover all possibilities.

- Prospective and Retrospective Studies:
  - Prospective studies collect data as events unfold, often through long-term observation.
  - Retrospective studies analyze past events using existing data.
  - Data sets might contain both prospectively and retrospectively collected variables.

- Implied Randomness and Observational Data:
  - Statistical methods rely on implied randomness in observational data collection.
  - Without random sampling, statistical methods lose reliability.

- Random Sampling Techniques:

  - Simple Random Sampling: Each case has an equal chance of being included; cases' inclusion does not impact others.
  - Stratified Sampling: Divides population into strata, similar cases grouped; then employs simple random sampling within each stratum.
  - Cluster Sampling: Breaks population into clusters, samples clusters, and performs simple random sampling within each cluster.

- Stratified Sampling:
  - Useful when cases within each stratum are similar with respect to the outcome of interest.
  - Enhances estimation stability for subpopulations within strata.
  - Requires more complex data analysis than simple random sampling.

- Cluster Sampling:
  - Similar to stratified sampling but doesn't necessitate sampling from every cluster.
  - Involves breaking population into clusters, sampling clusters, and performing simple random sampling within each cluster.

- Example Questions:
  - Sampling: Process of selecting a subset of individuals from a larger population for analysis. Example: Surveying salaries of MLB players.
  - Types of Sampling: Simple random, stratified, and cluster sampling.
  - Simulation: Replicating real-world conditions to predict outcomes. Example: Simulating evacuation plans for natural disasters.
  - Data Simulation Uses: Validating models, scenario testing, understanding variable impact.
  - Data Simulation Benefits: Modeling behavior, validation, visualization, strategy guidance.
  - Data Simulation Features: GUI, model building, scalability, analytics integration, data import/export.
  - Forms of Simulation Data: Approximating known conditions, experimenting with scenarios, climate projections, digital twins, etc.
  - Simulation and Predictive Analytics: Both require models but serve different purposes. Simulation models real-world conditions, while predictive analytics uses models for future insights.
  - Decision Tree vs. Machine Learning: Decision trees suitable for simple systems; machine learning handles complexity and large datasets better.
  - Two Types of Programmable Simulation Models: Mathematical models (e.g., compartmental models) and process models (e.g., agent-based models).