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

Read PPT

Observational Sampling Design

1. Introduction
- Observational studies collect data by monitoring events or phenomena without intervention.
- Experiments involve researchers assigning explanatory variables to subjects.
- Causal conclusions are often reliable from experiments but can be treacherous from observational studies.

2. Associations vs. Causation
- Observational studies generally show associations, not causation.
- Example: Sunscreen use and skin cancer association doesn't imply causation.

3. Confounding Variables
- Confounding variables are correlated with both explanatory and response variables.
- Sun exposure can confound the sunscreen-skin cancer association.
- It's difficult to account for all confounding variables.

4. Types of Observational Studies
- Prospective studies track individuals and events over time.
- Retrospective studies collect data after events have occurred.
- Example: The Nurses Health Study follows nurses over years to assess behavior's impact on cancer risk.

5. Randomness in Observational Studies
- Statistical methods rely on implied randomness.
- If observational data isn't collected randomly, statistical methods aren't reliable.

6. Random Sampling Techniques
- Simple random sampling: Each case has an equal chance of being included.
- Stratified sampling: Divide population into strata, sample within each stratum.
- Cluster sampling: Divide population into clusters, sample clusters, then within each cluster.
- Cluster sampling can be economical and useful for diverse clusters.

7. Cluster vs. Stratified Sampling
- Cluster sampling samples whole clusters, stratified sampling samples within strata.
- Cluster sampling might be more economical, while stratified sampling provides stable estimates within subpopulations.

8. Example: Choosing a Sampling Method
- Estimating malaria rate in a tropical portion of Indonesia.
- Simple random sampling might be expensive due to diverse villages.
- Stratified sampling could be challenging, so cluster sampling is a good choice.

These notes summarize the content about observational sampling design, explaining concepts like associations, confounding variables, types of observational studies, randomness, and different random sampling techniques. It also provides insights into when to use cluster vs. stratified sampling and illustrates a practical example of choosing a sampling method.
Observational Sampling Design Notes:

1. Observational Studies vs. Experiments:
- Observational studies collect data by monitoring events, while experiments involve researchers assigning variables.
- Causal conclusions from experiments are usually reasonable, but from observational data, they can be treacherous and not recommended.
- Observational studies generally show associations rather than causation.

2. Confounding Variable:
- Confounding variables (lurking variables) are correlated with both explanatory and response variables.
- Making causal conclusions from observational studies requires identifying and accounting for confounding variables.

3. Prospective and Retrospective Studies:
- Prospective studies collect data as events unfold; e.g., Nurses Health Study.
- Retrospective studies collect data after events have occurred, e.g., reviewing medical records.
- Data sets may contain both prospectively and retrospectively collected variables.

4. Implied Randomness and Sampling Techniques:
- Statistical methods rely on implied randomness in observational data.
- Three random sampling techniques: simple random, stratified, and cluster sampling.

5. Simple Random Sampling:
- Each case in the population has an equal chance of being included in the sample.
- Useful for unbiased representation in diverse populations.

6. Stratified Sampling:
- Divides population into groups (strata) with similar cases.
- Simple random sampling is then employed within each stratum.
- Useful when cases within strata are very similar regarding the outcome of interest.

7. Cluster Sampling:
- Population is divided into clusters, and a random sample of clusters is selected.
- Within each selected cluster, a simple random sample is taken.
- Economical when there's case-to-case variability within clusters but clusters are similar.

8. Cluster Sampling Example:

- If neighborhoods represent clusters, this method works best when neighborhoods are diverse but cases within each neighborhood are similar.

9. Choosing a Sampling Method:
- The choice of sampling method depends on the characteristics of the population and the research goals.

10. Example: Estimating Malaria Rate:
- Cluster sampling might be a good choice when there are similar villages and the goal is to test individuals for malaria.

These notes summarize key concepts from the provided text regarding observational sampling design, confounding variables, types of studies, and sampling techniques.

Reproducible Research

- Reproducible research involves publishing data analyses and scientific claims along with their data and software code, allowing others to verify findings and build upon them.
- The importance of reproducibility has grown due to complex data analyses with larger datasets and sophisticated computations.
- Reproducibility shifts focus from superficial details to the actual content of a data analysis, enhancing its usefulness.
- It makes analyses more valuable to others by providing access to the data and code used for the analysis.

Computational Reproducibility

- Replicating studies with new independent data is costly and methodologically challenging.
- Computational reproducibility, often called "reproducible research," is suggested to improve the assessment of scientific results' validity and rigor.
- Research is computationally reproducible when others can replicate study results using original data, code, and documentation.

Advantages of Reproducibility

- This approach mirrors the benefits of replicating studies with new data but minimizes the cost of collecting new data.
- While replicating studies remains the gold standard, reproducibility is considered a minimum standard for all scientists.

Principles of Reproducibility

- Researchers can adopt a simple three-part framework to make their current research more reproducible.
- These principles apply to researchers across various sub-disciplines.

Benefits of Reproducible Research

1. Researchers benefit from reproducible research by:
   - Ensuring consistent results upon multiple analyses.
   - Facilitating explanations of work to collaborators, supervisors, and reviewers.
   - Enabling quick and efficient supplementary analyses by collaborators.

2. Reproducible research enables easy modification of analyses and figures:
   - Responding to requests from supervisors, collaborators, and reviewers.
   - Saving significant time by updating figures through code changes.

3. Reproducible research simplifies reconfiguration of previous research tasks:
   - Simplifying subsequent projects requiring similar tasks.
   - Enhancing efficiency in iterative research processes.

4. Conducting reproducible research demonstrates rigor, trustworthiness, and transparency:
   - Increases the quality and speed of peer review.
   - Reviewers can directly access analytical processes in manuscripts.
   - Reviewers can cross-check code and methods, catching errors during peer review and reducing post-publication corrections.

Reproducible research benefits researchers, enhances collaboration, and ensures the reliability of scientific findings.

Why Do Reproducible Research?

Protects Against Accusations of Research Misconduct:
- Researchers who openly share code and data are less likely to be accused of research misconduct due to fraudulent practices.
- Fraudulent code and data would be evident to the research community.

Increases Paper Citation Rates:
- Reproducible research leads to higher citation rates for papers.
- Citations extend to code and data in addition to publications.
- Enhances the impact of research by making data and methods accessible.

Benefits the Research Community:
1. Facilitates Learning from Others' Work:
   - Allows researchers to access code and data, aiding in learning complex techniques.
   - Beginners can benefit from experienced researchers' code to perform rigorous analyses.

2. Saves Time and Effort for Experienced Researchers:
   - Experienced researchers can modify existing code more efficiently than writing from scratch.
   - Sharing code accelerates similar analyses for seasoned researchers.

3. Enables Understanding and Reproduction of Work:
   - Others can perform follow-up studies to strengthen evidence.
   - Promotes compatibility and consistency among similar studies.
   - Supports meta-analyses for generalizing and contextualizing findings.

4. Helps Identify and Correct Mistakes:
   - Open access to code and data encourages critical analysis.
   - Co-authors, reviewers, and other scientists can identify and rectify mistakes.
   - Prevents mistakes from accumulating over time.

Barriers to Reproducible Research:
- Complexity:
  - Specialized knowledge and tools required for certain analyses.
  - High-performance computing clusters with various programming languages.
  - Proprietary software like SAS or ArcGIS with expensive licenses.

- Technological Change:
  - Rapidly evolving technologies and tools complicate reproducibility.
  - New tools may not be widely available or understood.

- Human Error:
  - Mistakes can occur in scientific research.
  - Open access allows collaborators, reviewers, and others to catch errors early.

- Intellectual Property Concerns:
  - Fear of compromising intellectual property rights may hinder open sharing.
  - Protocols and norms can address these concerns and encourage openness.

Addressing Barriers:
- Complexity:
  - Citations and detailed annotations can reduce knowledge barriers.
  - Thoroughly annotated code and extensive documentation can enhance accessibility.

- Technological Change:
  - Researchers can actively work to bridge the technology gap by providing resources and tutorials.

- Human Error:
  - Open access and collaborative review help identify and correct mistakes.

- Intellectual Property Concerns:
  - Proper protocols and norms can balance openness with intellectual property rights.

Reproducible research benefits researchers, the scientific community, and the quality and reliability of scientific findings. Overcoming barriers through accessible resources and collaborative efforts is essential for fostering reproducibility.

Barriers to Reproducible Research

Technological Change:
- Hardware and software used for data analysis evolve rapidly.
- Research conducted with outdated tools becomes less reproducible over time.
- For instance, research from previous decades may require entirely new tools for replication today.
- Even minor updates in software can impact the reproducibility of a project.

Mitigation Through Established Tools:
- Use well-documented versions of software tools.
- Careful documentation of software versions is essential.

Human Error:
- Researchers make mistakes in documenting procedures and analyses.
- Incomplete descriptions and documentation can lead to inaccuracies.
- Critical data might be omitted initially but become vital later.

Documentation as a Safeguard:
- Detailed documentation guards against errors and incomplete analyses.
- Record data collection details, decisions, and labeling conventions.
- Data wrangling errors can be mitigated through multiple data backups and thorough documentation.

Intellectual Property Rights:
- Researchers may hesitate to share data and code due to misuse or unethical use.
- Sharing data without proper citation can lead to misinterpretations.
- Researchers might withhold data to protect their future analyses.

Balancing Openness and Protection:
- Emerging tools allow sharing while preserving control and credit.
- Open data sharing is a contentious aspect of reproducible research.

Framework for Reproducible Research

Before Data Analysis: Data Storage and Organization:
- Plan for reproducibility from the start with effective data management.
- Data should be backed up at every stage and stored in multiple locations.
- Backups should include raw and clean analysis-ready data.
- Keep paper copies of data sheets paired with digital datasets.
- Use portable, non-proprietary formats for digital data.

Addressing Technological Change:
- Use well-documented versions of software tools.
- Careful documentation of software versions is essential.

Human Error:
- Thorough documentation of processes guards against errors and incomplete analyses.

Intellectual Property Concerns:
- Emerging tools offer data sharing while safeguarding ownership and credit.

Framework for Conducting Reproducible Research

During Analysis: Best Coding Practices:

- Tidy Data Format: Transform data into a "tidy" format for cleaning and standardization. Tidy data are organized in long format, with consistent structure and informative headers.

- Metadata: Store metadata explaining data cleaning and variable meanings along with the data. Metadata enhances data interpretability and should include data collection details, variable meanings, and coding explanations.

- Organized File Structure: Organize files with informative names and directories. Consistent naming protocols for files and directories enhance searchability and accessibility.

- Version Control: Use version control to document project history and changes. This aids in tracking updates and provides snapshots of data and code.

During Analysis: Coding Practices:

- Use coding scripts for data wrangling and analysis for documentation and repeatability.
- Thoroughly annotate analytical code with comments for clarity and metadata.
- Follow consistent coding styles for readability.
- Automate repetitive tasks using functions and loops.
- Use parameters at the beginning of a script to allow easy adaptation to new data.

Mitigating Technological Change:
- Use established software versions and document dependencies.
- Consider using software containers for reproducibility.

After Analysis: Finalizing and Sharing Results:

- Share input data, scripts, program versions, parameters, and intermediate results publicly.
- Create figures and tables directly from code for dynamic, reproducible documents.
- Use tools like LaTeX for creating dynamic presentations.

Sharing and Archiving Results:

- Automation with Make: Use GNU Make to automate and coordinate command-line processes, making data wrangling, analysis, and document creation a streamlined process.

- Sharing Research: Currently, data and code for replicating research are often found in journal article supplementary materials. Some journals are experimenting with embedding data and code in articles. Authors can also post preprints on preprint servers or postprints on postprint servers to increase access to publications.

- Use of Data Repositories: Data archiving in online repositories is becoming more popular due to technology improvements, large-scale data sets, and encouragement from publishers and funding organizations. Repositories collect and store data for analysis, sharing, and reporting. Researchers can find appropriate repositories through journal recommendations.

- Research Compendia: Archiving data, code, software, and research products together forms a research compendium. These compendia provide a standardized way to organize and share research materials, making it easier for other researchers to reproduce and extend the research.

Three-Step Framework and Check-list Guide for Reproducible Research: This section provides a concise summary of the three-step framework for conducting reproducible research and emphasizes the importance of adopting these practices for improved research transparency and reliability.

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).

Descriptive Statistics Notes:

1. Introduction to Descriptive Statistics:
- Descriptive statistics summarize and organize characteristics of a data set.
- A data set consists of responses or observations from a sample or entire population.
- In quantitative research, the first step is describing the characteristics of the responses, like averages or relationships between variables.
- Inferential statistics come next, helping determine if data confirms hypotheses and can be generalized.

2. Types of Descriptive Statistics:
- Three main types: distribution, central tendency, and variability or dispersion.
- Distribution: Frequency of each value.
- Central Tendency: Averages of values.
- Variability/Dispersion: Spread of values.

3. Research Example:
- Studying leisure activity popularity by gender.
- Survey about past-year activities: library, movie theater, national park.
- Descriptive stats reveal activity frequency, averages, and spread.

4. Frequency Distribution:
- Data set consists of values, summarized in frequency distribution.
- Tabulate or graph frequency of each possible value of a variable.
- Example: Gender - Male: 182, Female: 235, Other: 27.

5. Measures of Central Tendency:
- Measures the center or average of a data set.
- Mean, median, and mode are common ways to find average.
- Example calculation using first 6 survey responses:
  - Mean = (15 + 3 + 12 + 0 + 24 + 3) / 6 = 9.5

6. Measures of Variability:
- Describes spread in response values.

- Range, standard deviation, and variance capture different aspects.
- Example calculation of standard deviation:
  - Steps include finding deviations, squaring them, summing, and taking the square root.
  - Standard deviation = 9.18.

7. Univariate Descriptive Statistics:
- Focuses on one variable at a time.
- Use multiple measures for distribution, central tendency, and spread.
- Example:
  - Visits to the library: Mean = 9.5, Median = 7.5, Mode = 3, SD = 9.18, Variance = 84.3, Range = 24.

8. Bivariate Descriptive Statistics:
- Explores relationships between two variables.
- Bivariate analyzes frequency, variability, and central tendency.
- Contingency tables and scatter plots help understand relationships.

9. Contingency Table:
- Intersection of two variables.
- Independent variable (e.g., gender) on vertical axis, dependent on horizontal.
- Percentages make interpretation easier.

10. Scatter Plots:
- Visualizes relationship between two or three variables.
- Data points represented on a chart.
- Used to assess correlations and perform regression tests.

These notes cover the concepts of descriptive statistics, different types of statistics, their calculations, and practical examples of their applications.

Introduction to Computational Data Analytics - Notes:

Computational Data Analytics:
- Field for specialization in data science (ML, deep learning, natural language, AI, etc.) building on interdisciplinary core curriculum.
- Computational thinking examples: chess strategy, map reading, task organization.
- Steps of Computational Thinking: Abstraction, Automation, Analysis.
- Principals of Computational Thinking: Decomposition, pattern recognition, abstraction, algorithms.
- Computational thinking benefits: Real-world problem solving, breaking down complex problems.

Computational Skills:
- Ability to perform basic arithmetic accurately and quickly using mental methods, calculators, etc.

Types of Computation:
- Models of computation: Sequential, functional, concurrent.

- Purpose of Computational: Intelligent health data analysis for disease treatment guidance.

Computational Analytics:
- Uses algorithms for pattern identification, anomaly detection, hypothesis testing, model creation, uncertainty quantification.
- Computational Data Science: Combines statistics, computer science, math, ML for trend identification, prediction, problem-solving.
- Uses algorithms, data structures for storage, manipulation, visualization, learning from large data sets.

Data Analytics:
- Examining data sets to discover trends, draw conclusions.
- Increasing use of specialized systems and software for data analytics.

Introduction to R Programming:
- Open-source language for statistical software, data analysis.
- Widely used, supports Windows, Linux, macOS.
- Command-line interface.

Why R Programming Language?:
- Leading tool for ML, statistics, data analysis.
- Platform-independent, open-source, integrates with other languages.
- Growing user community, high demand in Data Science job market.

Features of R Programming Language:
- Statistical Features: Basic statistics (mean, mode, median), static graphics, probability distributions, data analysis.
- Programming Features: Abundance of packages (CRAN), distributed computing.

Programming in R:
- Similar syntax to other languages, easy learning.
- Write .r programs, run using "R filename.r" command.

Advantages of R:
- Comprehensive statistical analysis package.
- Open-source, cross-platform, active community.

Disadvantages of R:
- Some package quality issues.
- Memory management challenges.
- Limited formal support.

Applications of R:
- Data Science, quantitative analysis, finance, tech giants like Google, Facebook.

R and Data Science:
- R and Python important in data science.
- Data science: Identify, represent, extract meaningful data for business logic.
- Data scientists use ML, statistics for analysis and decision-making.

Tools for Data Science:
- R, Python, SQL, SAS, Tableau, MATLAB, with R and Python being popular.
- Choosing between R and Python can be confusing for newcomers.

R vs Python in Data Science:
- R has advanced statistical techniques, Python covers common techniques.
- R packages cover various domains, Python's packages like Scikit-learn and Pandas are popular.
- R excels in data visualization, Python better for web development.
- Python better for deep learning and neural networks.
- R has abundant packages but may require more specialization, Python has fewer main packages but easier to use for common tasks.

Read PPT