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

READ PPT, DOC NOTES, LECT NOTES

PYQs AND UTs CHEATSHEET

UT1

Certainly! Here's a cheatsheet summarizing the answers to the questions provided:

1. Computational Data Analytics and Computational Thinking
   - Computational Data Analytics: Specialization in data science using algorithms for pattern recognition, hypothesis testing, modeling, and uncertainty quantification.
   - Computational Thinking Steps: Abstraction, Automation, Analysis, and Evaluation for problem-solving.

2. R-Programming Language
   - R is an open-source programming language for statistical analysis and data science.
   - Used for machine learning, statistics, and data analysis.
   - Statistical Features: Object creation, cross-platform compatibility, open-source nature, integration with other languages.

3. Measures of Central Tendency and Car Data
   - Measures of Central Tendency: Mean, Median, Mode.
   - Mean of Horsepower: Cannot be calculated with the given data.
   - Center Value of MPG: Median (22.8) as it represents the typical value.

4. Observational Sample Design
   - Observational Sampling Design: Collecting data without assigning primary explanatory variables.
   - Prospective Studies: Follow individuals over time.
   - Retrospective Studies: Collect data by looking back at past events.

5. Three Random Sampling Methods
   - Simple Random Sampling: Equal chance for each individual.
   - Stratified Sampling: Dividing the population into subgroups and sampling in proportion.
   - Cluster Sampling: Dividing into clusters and sampling clusters with all individuals in the selected clusters.

6. Barriers to Reproducible Research
   - Complexity: Specialized knowledge and tools.
   - Technological Change: Evolving hardware and software.
   - Human Error: Unintentional data errors.
   - Intellectual Property Rights: Concerns over data sharing.

7. Three-Step Framework for Conducting Reproducible Research
   - Before Data Analysis: Prepare data sources, collection methods, and preprocessing steps.
   - During Analysis: Use scripts or code for transparent analysis.
   - After Analysis: Share data, code, program versions, parameters, and intermediate results for accessibility and reproducibility.

Feel free to use this cheatsheet for quick reference.

UT2

Certainly, here's a cheatsheet summarizing the key points from the previous answers:

1. Diagnostic Analytics:
   - Examines past data to identify causes of events or outcomes.
   - Helps understand why certain events occurred and what factors contributed to them.

2. Working of Diagnostic Analytics:
   - Analyzes historical data to uncover patterns and causal relationships.
   - Utilizes algorithms, statistical techniques, and data visualization.

3. Probability Density Estimation:
   - Estimates the probability density function (PDF) of a random variable.
   - Kernel Density Estimation (KDE) is a common method.
   - Used to understand the likelihood of observing specific values.

4. Likelihood, Frequentist Approach, and Machine Learning:
   - Likelihood approach estimates parameters based on data.
   - Frequentist approach estimates parameters based on observed data frequencies.
   - Both are used in machine learning for modeling and inference.

5. Linear Regression Assumptions and Violations:
   - Assumptions: Linearity, Independence, Homoscedasticity, Normality.
   - Violations can be addressed through data transformation, robust regression, or more data collection.

6. Regularization in Linear Regression:
   - Prevents overfitting by adding a penalty term to the loss function.
   - L1 (Lasso) and L2 (Ridge) are common regularization techniques.
   - L1 encourages sparsity, while L2 shrinks coefficients.

7. Linear Regression for Data: [6, 8, 5, 10]
   - A linear regression model can be built to predict a target variable based on the given data points.
   - The model estimates the linear relationship between the independent variable (X) and the dependent variable (Y).

Use this cheatsheet for quick reference and study.

PYQ 8-12-2022

1. Data Structures in R:
   - Vectors: 1D arrays for the same data type.
   - Matrices: 2D arrays for the same data type.
   - Arrays: Multi-dimensional arrays for the same data type.
   - Lists: Collections of different data types.
   - Data Frames: 2D structures for different data types.

2. Top 3 Data Science Visualization Tools:
   - R: Packages like ggplot2, ggvis, and lattice.
   - Python: Bokeh and Matplotlib.
   - Tableau: User-friendly, interactive visualization tool.

3. Mean, Median, and Mode:
   - Mean: Sum of scores divided by the number of scores.
   - Median: Middle value when scores are arranged.
   - Mode: Most frequently occurring score.

4. Definitions:
   - Population: Entire group of interest.
   - Sample: Subset of the population.
   - Sampling: Process of selecting a sample.
   - Null Hypothesis: Assumes no significant difference.
   - Alternative Hypothesis: Assumes a significant difference.

5. Data Import:
   - Process of bringing external data into a computational environment.
   - Used in data simulation to make simulations more realistic.

6. Non-Probability Sampling:
   - Judgment Sampling: Selection based on domain knowledge.
   - Volunteer Sampling: Data from voluntary participants.
   - Convenience Sampling: Selecting readily available subjects.
   - Snowball Sampling: Participants recruit additional participants.

7. Model Selection:
   - Cross-validation: Assessing model performance on subsets of data.
   - Information criteria: Measures like AIC and BIC to compare models.

8. Data Plotting Types:
   - Scatter Plot: Visualizes relationships between two variables.
   - Line Plot: Shows trends or changes over time.
   - Bar Plot: Compares categorical data.
   - Histogram: Displays data distribution.

9. Power Analysis:
   - Determines sample size for detecting significant effects.
   - Considers effect size, significance level, and desired power level.

10. Linear Regression:
    - Models the relationship between two variables.
    - Least square regression line: $y = ax + b$.
    - Used for prediction and understanding relationships.

11. Data Analytics Life Cycle:
    - Data collection, data preparation, data analysis, data visualization.
    - Collect, clean, analyze, and present data.

12. Bayesian Inference:
    - Updates beliefs or knowledge based on new evidence.
    - Uses Bayes' theorem for posterior probability calculation.
    - Incorporates prior knowledge into analysis.

13. Challenges of Big Data Analytics:
   - Data volume, data sampling, data manipulation, data quality.
   - Privacy and security concerns.

14. Simple Random Sampling:
   - Selects a sample where each case has an equal chance.
   - Ensures unbiased representation from the population.

15. Measures of Central Tendency:
   - Mean: Sum divided by the number of values.
   - Median: Middle value when data is sorted.
   - Mode: Most frequently occurring value.

You can use these key points to create a concise cheatsheet for reference.

Certainly, here are cheatsheets for each of the answers provided for the questions in your previous requests:
PYQ 6-5-2023
Q: 1

a) Data Structures in R
   - Vectors
   - Matrices
   - Data Frames
   - Lists
   - Factors
   Cheatsheet:
   - Vectors: 1D arrays of the same data type.
   - Matrices: 2D arrays of the same data type.
   - Data Frames: Tabular data with columns of different data types.
   - Lists: Versatile data structures for mixed data types.
   - Factors: Represent categorical data.

b) Top 3 Data Science Visualization Tools
   - R
   - Python
   - Tableau
   Cheatsheet:
   - R: ggplot2, ggvis, lattice.
   - Python: Bokeh, Matplotlib.
   - Tableau: Interactive and dynamic visualizations.

c) Aspects of Data Visualization in Data Science
   - Data Exploration
   - Communication of Insights
   - Interactive Visualizations
   - Storytelling
   Cheatsheet:
   - Data Exploration: Identifying patterns and trends.

- Communication: Clear presentation of insights.
- Interactivity: Exploring data in-depth.
- Storytelling: Creating engaging data narratives.

d) Importance of Statistical Power
- Detecting True Effects
- Reducing Type II Errors
- Sample Size Determination
- Generalizability of Results

Cheatsheet:
- Detecting Effects: High power for true effects.
- Reducing Type II Errors: Avoiding false negatives.
- Sample Size: Determine sample size.
- Generalizability: Apply results broadly.

e) Probability Sampling vs. Non-Probability Sampling
- Probability Sampling: Known and non-zero probabilities.
- Non-Probability Sampling: No known probabilities.

Cheatsheet:
- Probability Sampling: Representative, controlled.
- Non-Probability Sampling: Potential bias, less control.

Q: 2

a) Data Plotting Types and Significance
- Scatter Plot
- Bar Chart
- Line Graph
- Histogram

Cheatsheet:
- Scatter Plot: Relationship between two variables.
- Bar Chart: Categorical data comparisons.
- Line Graph: Trends over time.
- Histogram: Distribution of continuous data.

b) Mean, Median, Mode of Test Scores
- Mean: 73.33
- Median: 77
- Mode: 79

Cheatsheet:
- Mean: Sum divided by the count.
- Median: Middle value in ordered data.
- Mode: Most frequent value.

Q: 3

a) Advantages and Disadvantages of Non-Probability Sampling
- Advantages: Cost-effective, quick, initial validation.
- Disadvantages: Non-representative sample, lack of control, potential bias.

Cheatsheet:
- Advantages: Cost, speed, validation.
- Disadvantages: Representativeness, control, bias.

b) Definitions
- Priori Power Analysis: Sample size estimation before data collection.
- z-score: Measure of data point's deviation from mean.
- Data Import: Bringing external data into analysis tools.
- Post-hoc Power Analysis: Calculating power after data analysis.
- Population: Entire group under study.
- Sample: Subset drawn for analysis.
- Sampling: Process of selecting a sample.

Cheatsheet:
- Priori Power: Estimate sample size in advance.
- z-score: Deviation from mean.
- Data Import: Bringing data into analysis tools.
- Post-hoc Power: Assess power after analysis.
- Population: Whole group of interest.
- Sample: Subset for analysis.
- Sampling: Selecting a sample.

Q: 4

a) Reproducibility and How to Achieve It
- Reproducibility Importance: Strengthens evidence, enhances compatibility, supports meta-analysis.
- Achieving Reproducibility: Document work, share data and code, strive for rigor.

Cheatsheet:
- Importance: Evidence, compatibility, meta-analysis.
- Achieving: Documentation, data/code sharing, rigor.

b) Frequentism
- Definition: Frequentism focuses on event frequency as a measure of probability.
- Justification: Probabilities are interpreted as long-run frequencies.

Cheatsheet:
- Definition: Frequency-based probability.
- Justification: Probabilities as long-run frequencies.

Q: 5

a) Bayesian Inference
- Definition: Bayesian Inference updates beliefs based on evidence.

Cheatsheet:
- Definition: Updating beliefs with data.

b) Data Simulation
- Definition: Data simulation creates virtual representations.

Cheatsheet:
- Definition: Creating virtual data models.

Q: 6

a) Linear Regression
- Definition: Linear regression models the relationship between two variables.

Cheatsheet:
- Definition: Modeling relationships with a linear equation.

b) Power Analysis
- Benefits: Estimate sample size, design experiments, evaluate reliability.
- Challenges: Accurate estimation, statistical assumptions, resource trade-offs, correct specification.

Cheatsheet:
- Benefits: Sample size, experiment design, reliability assessment.
- Challenges: Estimation, assumptions, resources, specification.

Q: 7

a) Simple Random Sampling
- Definition: Selecting a sample with equal probability for each case.

Cheatsheet:
- Definition: Equal probability sample selection.

b) Multi-stage Sampling
- Definition: Dividing the population into clusters and subsets for sampling.

Cheatsheet:
- Definition: Population division into clusters and subsets.

c) Quota Sampling
- Definition: Selecting participants based on specific quotas.

Cheatsheet:
- Definition: Selection based on specific quotas.

UT1 NOTES AND THEIR PPT NOTES

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

PPT 1
Introduction to Computational Data Analytics Cheat Sheet:

Computational Data Analytics:
- Field for depth and specialization in data science: ML, deep learning, natural language, AI, visualization, databases, high-performance computing, etc.
- Examples of computational thinking: chess strategy, map reading, task organization.
- Steps of Computational Thinking: Abstraction, Automation, Analysis.
- Principals of Computational Thinking: Decomposition, pattern recognition, abstraction, algorithms.
- Benefits of computational thinking: Real-world problem solving, simplifying complex problems.

Computational Skills:
- Ability to perform basic arithmetic quickly and accurately using mental methods, paper-and-pencil, or calculator.

Types of Computation:
- Models of computation: Sequential models, functional models, concurrent models.
- Purpose of Computational: Intelligent health information analysis for disease treatment guidance.

Computational Analytics:
- Enables scientific discovery through algorithms identifying patterns, anomalies, testing hypotheses, creating models.
- 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 datasets.

Data Analytics:
- Process of examining data sets to find trends and draw conclusions.
- Increasingly aided by specialized systems and software.

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

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

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

Programming in R:
- Similar syntax to other languages, easy to learn and code.
- Write and save programs with .r extension, run with "R filename.r" command.

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

Disadvantages of R:
- Some packages may have lower quality.
- Memory management challenges.
- Limited formal support.

Applications of R:
- Data Science, quantitative analysis, finance, tech giants like Google, Facebook, etc.
R and Data Science:
- R and Python play major roles in data science.
- Data science involves identifying, representing, extracting meaningful information for business logic.

Tools for Data Science:
- R, Python, SQL, SAS, Tableau, MATLAB.
- R and Python are most used.
R vs Python in Data Science:
- Specialities: R has advanced statistical techniques, Python is good for web development.
- Functionalities: R has inbuilt data analysis functions, Python relies on packages.
- Domains: R excels in data visualization, Python in deep learning.
- Packages: R has numerous packages, Python relies on Scikit-learn, Pandas.
Remember, this cheat sheet provides a concise overview of the key concepts covered in the introduction to Computational Data Analytics, including aspects of Computational Thinking, R Programming, and the comparison between R and Python for Data Science.

SYLLABUS CHEATSHEET

MODULE 1:

Sure! Here's a cheatsheet covering the topics you mentioned: Introduction to R Computing language, Reproducible Research in data science, Sampling and Simulation, Descriptive statistics, and creation of good observational sampling designs.

Introduction to R Computing Language:

1. R is a programming language and software environment for statistical computing and graphics.
2. Use the R console or an Integrated Development Environment (IDE) like RStudio to interact with R.
3. R uses functions and packages to perform specific tasks. Install packages using the `install.packages()` function and load them using `library()`.

Reproducible Research in Data Science:

1. Organize your project by creating separate folders for data, code, figures, and reports.
2. Use version control systems like Git to track changes in your code and collaborate with others.
3. Document your code using comments and markdown files to provide context and explanations.
4. Use RMarkdown or Jupyter Notebooks to combine code, visualizations, and text in a single document.
5. Set a random seed using `set.seed()` to ensure reproducibility in random processes.

Sampling and Simulation:

1. Use the `sample()` function in R to randomly sample from a population.
2. Specify the sample size and the population from which to sample.
3. For simulation studies, use loops (`for` or `while`) to repeatedly perform a task with different parameters or random inputs.
4. Store the results of each iteration in a data structure (e.g., vector, matrix, or list).
5. Visualize the results using plots or summary statistics to analyze the simulation outcomes.

Descriptive Statistics:

1. Use the `summary()` function to get a summary of the main statistics (minimum, 1st quartile, median, mean, 3rd quartile, maximum) for a numeric variable.
2. Calculate the mean using `mean()` and the median using `median()`.
3. Use `sd()` to compute the standard deviation and `var()` for the variance.
4. Obtain the correlation coefficient between two variables using `cor()`.
5. Create box plots, histograms, or scatter plots to visualize the distribution and relationships of variables.

Creation of Good Observational Sampling Designs:

1. Clearly define the target population and the variables of interest.
2. Ensure the sample is representative of the population by using random sampling techniques.
3. Use stratified sampling when the population can be divided into homogeneous subgroups.
4. Consider the sample size needed to achieve sufficient statistical power.
5. Document the sampling process, including the sampling method used and any biases that may be present.

Remember, this cheatsheet provides a brief overview of the topics mentioned. Further exploration and learningare encouraged to gain a deeper understanding of each area.

MODULE 2:

Certainly! Here's a cheatsheet covering the topics you mentioned: Data visualization, Data import and visualization, Introduction to various plots, Frequentist Hypothesis Testing, Z-Tests, and Power Analysis.

Data Import and Visualization:

1. Use the `read.csv()` function to import data from a CSV file into R.
2. Explore the structure of your data using functions like `str()` and `head()`.
3. Clean and preprocess the data by handling missing values, transforming variables, and filtering unwanted observations.
4. Visualize data using packages like ggplot2 or base R's plotting functions (`plot()`, `hist()`, etc.).
5. Customize plots by adding titles, labels, legends, colors, and themes.

Introduction to Various Plots:
1. Scatter Plot: Use `plot()` with two numeric variables to display the relationship between them.
2. Bar Plot: Use `barplot()` or `geom_bar()` in ggplot2 to represent categorical data as bars.
3. Histogram: Use `hist()` or `geom_histogram()` to visualize the distribution of a numeric variable.
4. Box Plot: Use `boxplot()` or `geom_boxplot()` to display the distribution of a numeric variable across differentcategories.
5. Line Plot: Use `plot()` or `geom_line()` to show the trend or change in a numeric variable over time or another continuous variable.
6. Heatmap: Use `heatmap()` or `geom_tile()` to represent data in a matrix-like form using colors.
7. Pie Chart: Use `pie()` or `geom_bar()` with a polar coordinate system to display proportions of a categorical variable.

Frequentist Hypothesis Testing:
1. Formulate the null hypothesis (H0) and alternative hypothesis (Ha) based on the research question.
2. Choose an appropriate test statistic based on the data and research question (e.g., mean, proportion, difference in means, etc.).
3. Set the significance level (α), typically 0.05, to determine the threshold for rejecting the null hypothesis.
4. Calculate the test statistic (e.g., z-score) using the sample data and relevant formulas.
5. Compare the test statistic to the critical value(s) from the appropriate distribution (e.g., standard normal distribution for z-tests) to make a decision about the null hypothesis.
6. Report the p-value, which represents the probability of obtaining results as extreme or more extreme than what was observed, assuming the null hypothesis is true.

Z-Tests:
1. Z-Test for a Population Mean: Use when you have a large sample size ($n > 30$) or know the population standard deviation.
2. Calculate the z-score using the formula: $z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$, where $\bar{x}$ is the sample mean, μ is the population mean, σ is the population standard deviation, and n is the sample size.
3. Compare the z-score to the critical value(s) from the standard normal distribution or calculate the p-value to make a decision.

Power Analysis:
1. Power analysis helps determine the sample size needed to detect a specific effect size with a desired level of statistical power.
2. Specify the effect size (difference between groups or association strength), significance level (α), and desired power ($1 - \beta$).
3. Use power analysis functions or online calculators specific to the statistical test you plan to conduct (e.g., t-test, ANOVA, correlation).
4. Adjust the sample size, effect size, or significance level to achieve the desired level of power.
Remember, this cheatsheet provides a brief overview of the topics mentioned. Further exploration and learningare encouraged to gain a deeper understanding of each area.

MODULE 3:
Certainly! Here's a cheatsheet covering the topics you mentioned: Linear regression, diagnostics, visualization,Likelihoodist Inference, fitting a line with likelihood, and model selection with one predictor.

Linear Regression:
1. Linear regression models the relationship between a dependent variable (response) and one or more independent variables (predictors).
2. Fit a linear regression model using the `lm()` function in R: `lm(y ~ x1 + x2, data = df)`, where `y` is the dependent variable and `x1`, `x2` are the predictors.
3. Extract the model coefficients using `coef()`: `coef(model)`.
4. Obtain the predicted values using `predict()`: `predict(model, newdata = df)`.
5. Evaluate the model's goodness of fit using metrics like R-squared (`summary(model)$r.squared`), adjusted R-squared (`summary(model)$adj.r.squared`), and root mean squared error (RMSE).

Diagnostics and Visualization:
1. Plot the residuals against the fitted values using `plot(model, which = 1)`.
2. Check for heteroscedasticity by plotting the standardized residuals against the fitted values using `plot(model, which = 3)`.
3. Use a normal probability plot (`plot(model, which = 2)`) to assess the normality of residuals.
4. Plot the Cook's distance to identify influential observations using `plot(model, which = 4)`.
5. Use diagnostic plots like residual vs. predictor variables or leverage plots to identify influential points or potential problems.

Likelihoodist Inference:
1. Likelihoodist inference is based on the likelihood function, which represents the probability of observing the data given the model parameters.
2. Fit a likelihood-based model using the `glm()` function in R: `glm(y ~ x1 + x2, data = df, family = gaussian)`, where `gaussian` specifies the distributional assumption.
3. Extract the model coefficients and their standard errors using `coef()` and `summary()`.
4. Perform hypothesis tests using likelihood ratio tests (`anova(model, test = "LRT")`), Wald tests (`summary(model)`), or score tests (`summary(model)$coefficients`).

Fitting a Line with Likelihood:
1. Fit a linear model using maximum likelihood estimation (MLE) by assuming the errors follow a specific distribution (e.g., Gaussian).
2. Use the `glm()` function with `family = gaussian` to fit the model: `glm(y ~ x, data = df, family = gaussian)`.
3. Extract the coefficients and their standard errors using `coef()` and `summary()`.
4. Evaluate the model using goodness-of-fit measures like deviance or Akaike Information Criterion (AIC).

Model Selection with One Predictor:
1. Fit multiple linear regression models with different predictor variables.
2. Compare models using goodness-of-fit measures like R-squared or adjusted R-squared.
3. Use the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) to compare models, where lower values indicate better fit.
4. Select the model with the highest R-squared or lowest AIC/BIC as the "best" model for prediction or inference.
Remember, this cheatsheet provides a brief overview of the topics mentioned. Further exploration and learningare encouraged to gain a deeper understanding of each area.

MODULE 4:
Sure! Here's a cheatsheet covering the topics you mentioned: Bayesian Inference, Fitting a line with Bayesian techniques, Multiple Regression and Interaction Effects, and Information Theoretic Approaches.

Bayesian Inference:
1. Bayesian inference is a framework for updating beliefs about unknown parameters using Bayes' theorem.
2. Specify a prior distribution representing your initial beliefs about the parameters.
3. Calculate the posterior distribution by combining the prior distribution with the likelihood function.
4. Summarize the posterior distribution using statistics like the mean, median, or credible intervals.
5. Markov Chain Monte Carlo (MCMC) methods, such as the Metropolis-Hastings algorithm or Gibbs sampling,are commonly used for Bayesian inference.

Fitting a Line with Bayesian Techniques:
1. Fit a linear regression model using Bayesian techniques by specifying prior distributions for the coefficients.
2. Use packages like 'rstan' or 'brms' in R to fit Bayesian linear regression models.
3. Specify the prior distribution for the coefficients using distributional assumptions such as normal, Student's t, or shrinkage priors.
4. Perform posterior inference by sampling from the posterior distribution using MCMC methods.
5. Visualize the posterior distribution of the coefficients and make inferences based on the credible intervals.
Multiple Regression and Interaction Effects:
1. Extend linear regression models to include multiple predictors.
2. Fit a multiple regression model using the `lm()` function in R: `lm(y ~ x1 + x2 + x3, data = df)`, where `y` is the dependent variable, and `x1`, `x2`, `x3` are the predictors.
3. Include interaction terms by multiplying the predictors: `lm(y ~ x1 + x2 + x1x2, data = df)`.
4. Interpret the regression coefficients as the change in the dependent variable associated with a one-unit change in the predictor, holding other predictors constant.
5. Assess the significance of the predictors and interaction terms using hypothesis tests or credible intervals from Bayesian models.

Information Theoretic Approaches:
1. Information theoretic approaches, such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), help compare and select among different models.
2. Calculate the AIC for a model using `AIC(model)`, where lower values indicate a better fit.
3. Calculate the BIC for a model using `BIC(model)`, which penalizes model complexity more than AIC.
4. Compare models using the differences in AIC or BIC values, with smaller differences indicating stronger evidence for a particular model.
5. Select the model with the lowest AIC or BIC as the "best" model, considering both fit and model complexity.
Remember, this cheatsheet provides a brief overview of the topics mentioned. Further exploration and learningare encouraged to gain a deeper understanding of each area.

UT2 NOTES AND THEIR PPT NOTES
Here is your cheatsheet on "Visualization in Computational Data Analytics," "Model Selection with One Predictor in Computational Data Analytics," "Linear Regression," "Assumptions of Linear Regression," and "Diagnostics Analytics":

Visualization in Computational Data Analytics:

1. Data Exploration:
   - Initial understanding of data's structure, distribution, and patterns.

- Essential for data exploration.

2. Pattern Recognition:
   - Reveals hidden patterns or trends.
   - Identifies anomalies or outliers.

3. Communication:
   - Powerful means of conveying complex information.
   - Aids in decision-making processes.

4. Comparisons:
   - Easy comparison of data across categories, time periods, or groups.
   - Simplifies drawing insights and conclusions.

5. Dimensionality Reduction:
   - Techniques like PCA and t-SNE for visualizing high-dimensional data in lower dimensions.

6. Interactive Visualizations:
   - Created using tools like D3.js and Tableau.
   - Enables dynamic data exploration.

7. Data Cleaning:
   - Highlights data quality issues.
   - Aids in preprocessing.

Model Selection with One Predictor in Computational Data Analytics:

1. Univariate Analysis:
   - Examines the relationship between a single predictor and the target variable.
   - Descriptive statistics and basic visualizations.

2. Correlation Analysis:
   - Assesses the correlation between the single predictor and the target variable.

3. Hypothesis Testing:
   - Utilizes tests like t-tests or ANOVA to determine the predictor's significance.

4. Feature Selection:
   - Evaluates the importance of the single predictor.
   - May exclude weak predictors.

5. Model Building:
   - If significant, build a simple regression model (e.g., linear regression).

6. Model Evaluation:
   - Assess model performance with metrics (e.g., R-squared, MAE).

7. Model Validation:
   - Use techniques like cross-validation to ensure model generalizability.

8. Model Comparison:
   - Compare the model with other potential predictors or models.

Linear Regression:

- Introduction:
  - Supervised machine learning model.
  - Establishes a linear relationship between input (independent) and output (dependent) variables.

- Equations:
  - Simple Linear Regression: `y = b0 + b1x`
  - Multiple Linear Regression: `y = b0 + b1x1 + b2x2 + ... + bnxn`

- Objective:
  - Find the best-fit line, minimizing error.

- Assumptions of Linear Regression:
  1. Linearity
  2. Normality
  3. Homoscedasticity
  4. Independence
  5. Error Terms Distribution
  6. No Autocorrelation

- Techniques:
  - Ordinary Least Squares (OLS), Gradient Descent, Regularization.

- Applications:
  - Marketing, finance, insurance, etc.

Assumptions of Linear Regression:

1. Linearity:
   - Dependent variable linearly related to independent variables.

2. Normality:
   - Both variables should be normally distributed.

3. Homoscedasticity:
   - Variance of error terms should be constant.

4. Independence/No Multicollinearity:
   - Independent variables uncorrelated, no multicollinearity.

5. Error Terms Distribution:
   - Error terms should be normally distributed.

6. No Autocorrelation:

- Error terms independent of each other.

Diagnostics Analytics:

- What Is Diagnostic Analytics?:
  - Explains "Why did this happen?"
  - Identifies causative factors in data.

- Importance:
  - Gain insights into factors affecting events.
  - Improve decision-making.

- Types of Analytics:
  - Descriptive, Predictive, Prescriptive, Diagnostic.

- How Does Diagnostic Analytics Work?:
  - Data drilling, data mining, correlation analysis.
  - Identify anomalies, gather data, establish causal connections.

- Process:
  1. Identify Anomalies
  2. Discovery
  3. Establish Causal Connections

- Use Cases:
  - Healthcare, retail, manufacturing, human resources.

- Benefits:
  - Understand reasons behind past events.
  - Informed decision-making.

- Drawbacks:
  - Focus on historical data.
  - Complement with predictive and prescriptive analytics.

Likelihood Frequentist:

- Introduction to Likelihood:
  - Likelihood vs. Probability.

- Maximum Likelihood Estimation (MLE):
  - Estimating model parameters from data.

- Models:
  - Formal representation of events or processes.

- Introduction to Maximum Likelihood Estimation for Machine Learning:
  - Solving density estimation problems.
  - MLE and likelihood function.

- Problem of Probability Density Estimation:
  - Estimating joint probability distribution for a dataset.
  - MLE and MAP approaches.

- Maximum Likelihood Estimation:
  - Optimization to maximize likelihood.
  - Using log-likelihood function.

- Relationship to Machine Learning:
  - Application in supervised and unsupervised learning.

- Fitting a Line using Likelihood:
  - Linear regression as an MLE problem.
  - Derivation and goal of MLE equation.

These cheatsheets cover key concepts in data analytics, visualization, model selection, linear regression, assumptions, diagnostics analytics, and likelihood frequentist.

Visualization in Computational Data Analytics:

1. Data Exploration: Visualization is a critical step in data exploration. It helps analysts gain an initial understanding of the data's structure, distribution, and patterns.

2. Pattern Recognition: Visualizing data can reveal hidden patterns or trends that might not be apparent from raw data. Graphs, charts, and plots make it easier to identify anomalies or outliers.

3. Communication: Visualizations serve as a powerful means of communication. They make it easier to convey complex information to non-technical stakeholders, helping in decision-making processes.

4. Comparisons: Visual representations allow for the easy comparison of data across different categories, time periods, or groups, making it simpler to draw insights and conclusions.

5. Dimensionality Reduction: Techniques like PCA (Principal Component Analysis) and t-SNE (t-distributed Stochastic Neighbor Embedding) are used for visualizing high-dimensional data in lower dimensions to uncover structures or clusters.

6. Interactive Visualizations: Interactive visualizations, created using tools like D3.js or Tableau, enable users to explore data dynamically, promoting deeper insights and understanding.

7. Data Cleaning: Visualizations can highlight data quality issues, such as missing values or inconsistencies, making it easier to address these problems during the preprocessing stage.

Model Selection with One Predictor in Computational Data Analytics:

1. Univariate Analysis: In cases where you have one predictor variable, univariate analysis involves examining the relationship between the single predictor and the target variable. This could be done using descriptive statistics and basic visualizations.

2. Correlation Analysis: Assess the correlation between the single predictor and the target variable to understand the strength and direction of the relationship.

3. Hypothesis Testing: Perform hypothesis tests like t-tests or ANOVA to determine if the single predictor significantly affects the target variable.

4. Feature Selection: Evaluate the importance of the single predictor in the context of your modeling goals. If it's a weak predictor, you might consider excluding it from your model.

5. Model Building: If the single predictor is deemed significant and relevant, you can build a simple regression model, like a linear regression, to predict the target variable using this predictor.

6. Model Evaluation: Assess the model's performance using appropriate metrics (e.g., R-squared, Mean Absolute Error, etc.) to determine how well the single predictor explains the variability in the target variable.

7. Model Validation: Use techniques like cross-validation to ensure the model's generalizability and robustness.

8. Model Comparison: If you have other potential predictors or models, compare the model built with this single predictor to models with additional predictors to determine which one performs better in terms of predictive accuracy and generalization.

These topics are related to Computational Data Analytics as they involve the process of analyzing and extracting insights from data, whether through visualization techniques for data exploration or model selection for predictive modeling, which is a key aspect of data analytics and machine learning.

Linear Regression PPT Notes
Certainly, here are detailed notes on Linear Regression:

Introduction
- Linear Regression is a supervised Machine Learning model that finds the best-fit linear line between independent and dependent variables, establishing a linear relationship.
- It assumes a linear relationship between input variables (independent variables 'x') and the output variable (dependent variable 'y').
- Simple Linear Regression is used for a single input variable, while Multiple Linear Regression is employed when there are multiple input variables.

Equations
- Simple Linear Regression Equation: `y = b0 + b1x`, where `b0` is the intercept, `b1` is the coefficient or slope, `x` is the independent variable, and `y` is the dependent variable.

- Multiple Linear Regression Equation: `y = b0 + b1x1 + b2x2 + b3x3 + ... + bnxn`, where `b0` is the intercept, `b1`, `b2`, `b3`, ..., `bn` are coefficients or slopes of independent variables `x1`, `x2`, `x3`, ..., `xn`, and `y` is the dependent variable.

Objective
- The primary goal of a Linear Regression model is to find the best-fit linear line with optimal values for the intercept and coefficients that minimize the error (the difference between actual and predicted values).

Mathematical Approach

- Residual/Error: Residual/Error represents the difference between actual values and predicted values.
- Sum of Residuals/Errors: The sum of the differences between actual and predicted values.
- Square of Sum of Residuals/Errors: The square of the sum of the differences between actual and predicted values.

Assumptions of Linear Regression
1. Linearity: The dependent variable should be linearly related to independent variables. This can be verified through scatter plots.

2. Normality: Both the independent and dependent variables should be normally distributed.

3. Homoscedasticity: The variance of error terms should be constant, meaning the spread of residuals should be consistent.

4. Independence/No Multicollinearity: Independent variables should be uncorrelated, with no significant multicollinearity.

5. Error Terms Distribution: Error terms should be normally distributed.

6. No Autocorrelation: Error terms should be independent of each other.

Dealing with Assumption Violations
- Violations of assumptions can lead to decreased model accuracy. Techniques for handling these violations include data transformations, feature selection, or regularization methods.

Evaluation Metrics for Regression Analysis
1. R-squared (Coefficient of Determination): Measures the proportion of the variance in the dependent variable explained by the independent variables.

2. Adjusted R-squared: Adjusts R-squared for the number of independent variables to provide a more accurate representation of model performance.

3. Mean Squared Error (MSE): The mean of the squared differences between actual and predicted values.

4. Root Mean Squared Error (RMSE): The square root of the MSE, penalizing larger errors.

Model Representation
- In Simple Linear Regression with one input variable and one output variable, the model takes the form: `Y(pred) = b0 + b1x`.

- In higher dimensions (with multiple input variables), the line becomes a plane or hyperplane.

Violations of Assumptions
- Violations of linearity, independence, homoscedasticity, and normality assumptions can lead to inaccurate model results and predictions.

Assumptions of Linear Regression:

Linear regression is a widely used statistical technique for modeling relationships between dependent and independent variables. To justify the use of linear regression for inference and prediction, several key assumptions must be satisfied:

1. Linearity and Additivity:
   - (i) The relationship between the dependent variable (Y) and each independent variable (X) is both linear and additive.
   - (a) The expected value of Y is a straight-line function of each independent variable while holding all other variables constant.
   - (b) The slope of this line remains constant and does not depend on the values of other independent variables.
   - (c) The effects of different independent variables on the expected value of Y are additive, meaning they do not interact in a multiplicative or nonlinear manner.

2. Statistical Independence of Errors:
   - (ii) The errors (residuals) are statistically independent of each other.
   - Specifically, there should be no correlation between consecutive errors in the case of time series data. Each observation's error is not influenced by the error of the previous observation.

3. Homoscedasticity (Constant Variance) of Errors:
   - (a) For time series data, the variance of errors should be constant over time. There should be no systematic change in the spread of errors as time progresses.
   - (b) For predictions, the variance of errors should be constant for all values of the independent variables. The spread of errors should not change as you make predictions.
   - (c) The variance of errors should not vary systematically with any of the independent variables. In other words, the errors should exhibit constant variance across the entire range of independent variable values.

4. Normality of the Error Distribution:
   - (iv) The errors (residuals) should follow a normal distribution.
   - This assumption implies that the errors are symmetrically distributed around zero and follow a bell-shaped curve.

Impact of Violating Assumptions:
If any of these assumptions are violated, meaning that there are nonlinear relationships between dependent and independent variables or the errors exhibit correlation, heteroscedasticity (varying error variance), or non-normality, the following consequences may occur:
- Forecasts: Predictions made by the regression model may be inaccurate.
- Confidence Intervals: Confidence intervals for parameter estimates may be unreliable.
- Scientific Insights: Conclusions drawn from the analysis may be inefficient, biased, or misleading.

Therefore, it is essential to check these assumptions when applying linear regression models and consider alternative methods or adjustments when they are not met. Violations of these assumptions can lead to unreliable results and hinder the interpretability of the model.

Techniques for Building a Linear Regression Model
1. Ordinary Least Squares (OLS): Minimizes the sum of squared differences between observed and predicted values to fit a multiple linear regression model.

2. Gradient Descent: Iteratively minimizes the error in the model by updating coefficients using a learning rate.

3. Regularization: Extensions of linear regression that reduce complexity and account for collinearity, including Lasso and Ridge Regression.

Applications of Linear Regression
- Linear regression can be applied in various fields such as marketing, finance, and insurance to evaluate trends, make forecasts, analyze marketing effectiveness, assess risk, and optimize decision-making processes.

Real-time Example
- Linear regression can be used to predict student grades based on the number of hours studied, with the objective of minimizing the prediction error.

These detailed notes cover the fundamental concepts, assumptions, techniques, evaluation metrics, and applications of Linear Regression.

Diagnostics PPT Notes
Diagnostics Analytics: Detailed Notes

What Is Diagnostic Analytics?

- Definition: Diagnostic analytics is a branch of advanced analytics that focuses on answering the question, "Why did this happen?" It involves examining data or content to understand the root causes of observed patterns and events.

- Techniques: Diagnostic analytics utilizes various techniques, including data drilling, data mining, and correlation analysis, to delve into data and identify causative factors.

- Additional Data: In some cases, to investigate the root causes of trends, companies may need to incorporate external data sources alongside internal data.

- Purpose: The primary purpose of diagnostic analytics is to help organizations gain insights into the factors driving their past events and make informed decisions to remedy issues and improve future outcomes.

Importance of Diagnostic Analytics

- Diagnostic analytics is vital for companies in gaining a comprehensive understanding of their business performance.

- It aids in discerning the influence of both internal and external factors on outcomes, helping companies make better-informed decisions.

- It is particularly valuable for identifying the reasons behind trends and events and enables companies to replicate success and rectify problems.

- For example, if a specific marketing campaign led to increased product sales, diagnostic analytics can uncover this and guide the allocation of more resources to similar campaigns.

Types of Analytics

- Diagnostic analytics is one of the four primary types of business analytics, alongside descriptive, predictive, and prescriptive analytics.

- Descriptive analytics focuses on summarizing and highlighting historical data trends to answer "What happened?"

- Predictive analytics looks into how future trends might unfold and their potential impact.

- Prescriptive analytics suggests actions to respond to future trends and improve business outcomes.

How Does Diagnostic Analytics Work?

- Diagnostic analytics employs techniques like data drilling, data mining, and correlation analysis to identify the causes of trends.

- Data drilling involves a deeper examination of specific aspects of the data to discover what is driving observed trends.

- Data mining searches for patterns and associations within data, revealing the most common factors linked to specific events.

- Correlation analysis assesses the strength of relationships between different variables in the data.

Process of Diagnostic Analytics

- The diagnostic analytics process typically comprises three stages:

  1. Identify Anomalies: Recognize trends or anomalies that require explanation, sometimes using statistical analysis to confirm their significance.

  2. Discovery: Gather data that can explain the anomalies, which may include external data sources alongside internal data.

  3. Establish Causal Connections: Use techniques like probability theory, regression analysis, filtering, and time-series data analytics to determine causal relationships among variables and uncover the root causes of anomalies.

Three Diagnostic Analytics Categories

- The diagnostic analytics process can be categorized into three stages: identifying anomalies, conducting data discovery, and establishing causal relationships.

Use Cases of Diagnostic Analytics

- Diagnostic analytics is applicable in various industries, including healthcare, retail, manufacturing, and human resources.

- It can be used to investigate the causes of trends such as revenue fluctuations, product popularity, employee turnover, and production bottlenecks.

Benefits of Diagnostic Analytics

- Diagnostic analytics helps companies understand the reasons behind past events, facilitating informed decision-making and a data-driven culture.

- It allows businesses to identify contributing factors that may not be immediately apparent, enabling more effective solutions and improvements.

Drawbacks of Diagnostic Analytics

- A limitation of diagnostic analytics is its focus on historical data; it does not provide insights into future events.

- It may require further investigation to establish definitive cause-and-effect relationships between variables.

- To address future trends, businesses should complement diagnostic analytics with predictive and prescriptive analytics.

Likelihood Frequentist PPT Notes
Likelihood Frequentist: Detailed Notes

Introduction to Likelihood
- Likelihood describes how to find the best distribution of the data for some feature or situation in the data given a certain value of some feature or situation.
- Probability describes how to find the chance of something given a sample distribution of data.

Maximum Likelihood Estimation (MLE)
- MLE is a frequentist approach for estimating the parameters of a model given some observed data.
- General approach:
  1. Observe some data.
  2. Write down a model for how the data was generated.
  3. Set the model parameters to values that maximize the likelihood of the parameters given the data.

Models
- A model is a formal representation of beliefs, assumptions, and simplifications surrounding an event or process.
- Example: Coin Flip
  - Factors to consider: the coin's properties, initial position, force exerted, angle of force, center of mass, gravity.
- Simplified models are often used for practicality even when the real world is complex.

Introduction to Maximum Likelihood Estimation for Machine Learning
- Density estimation is about estimating the probability distribution for a sample of observations.
- Maximum Likelihood Estimation is a common framework for solving density estimation problems.
- It involves defining a likelihood function to calculate the conditional probability of observing data given a probability distribution and distribution parameters.

Problem of Probability Density Estimation

- Probability density estimation involves estimating the joint probability distribution for a dataset.
- It's challenging when the sample is small and noisy.
- Two common approaches: Maximum a Posteriori (MAP) and Maximum Likelihood Estimation (MLE).

Maximum Likelihood Estimation
- MLE is an optimization problem to find parameters that maximize the likelihood function.
- Likelihood function is the conditional probability of observing the data given the model parameters.
- It's common to use the log-likelihood function to avoid numerical instability when multiplying many small probabilities.
- Minimizing the negative log-likelihood (NLL) is often preferred in optimization problems.

Relationship to Machine Learning
- MLE can be applied to supervised and unsupervised machine learning.
- MLE can be used to estimate conditional probabilities, e.g., predicting the output (y) given the input (X).
- Linear regression and logistic regression can be framed as MLE problems.
- MLE provides a consistent way to approach predictive modeling as an optimization problem.

Fitting a Line using Likelihood
- Linear regression can be framed as a maximum likelihood problem.
- The likelihood function is derived from assuming Gaussian noise.
- The goal is to maximize the product of the probabilities for each data point, which is equivalent to the NLL minimization.
- Linear regression provides a prediction that is the mean of a Gaussian distribution, which can be used to calculate prediction intervals.

Derivation of MLE Equation
- A brief derivation of the MLE equation for linear regression is provided.
- The likelihood function is derived using the assumption of normally distributed errors.
- The MLE equation for linear regression is presented.

These detailed notes cover the concepts of Likelihood Frequentist, Maximum Likelihood Estimation, the role of likelihood in machine learning, and its application in linear regression. The notes also include equations and explanations to help understand the concepts better.

READ PPT IMPORTANT BEFORE EXAM