Content Ref: Dr. Anita Morey Mam Presentations

PYQ 6-5-2023

Q: 1 Solve any three.
a) What are the different data structures in R? Briefly explain about them.
ANS.

Data Structures in R

R programming language provides several data structures that are used to store and manipulate data. Here are some of the main data structures in R:

1. Vectors: Vectors are one-dimensional arrays that can hold elements of the same data type. They can be created using the `c()` function and can store numeric, character, or logical values.
2. Matrices: Matrices are two-dimensional arrays that can store elements of the same data type. They can be created using the `matrix()` function and are useful for performing matrix operations and calculations.
3. Data Frames: Data frames are tabular data structures that can store different types of data. They are similar to matrices, but each column can have a different data type. Data frames are commonly used for data analysis and manipulation.
4. Lists: Lists are versatile data structures that can store elements of different data types. They can be created using the `list()` function and can contain vectors, matrices, data frames, or even other lists.
5. Factors: Factors are used to represent categorical data in R. They are created using the `factor()` function and are useful for statistical analysis and modeling.

These data structures in R provide flexibility and efficiency in handling and analyzing data, making R a powerful language for data manipulation and statistical analysis.

b) Give any Top 3 Data Science Visualization Tools.
ANS.
Top 3 Data Science Visualization Tools

1. R: R is a powerful tool for data visualization with packages like ggplot2, ggvis, and lattice, making it easier to create visually appealing plots and charts.
2. Python: Python is catching up in the field of data visualization with packages like Bokeh and Matplotlib. Although it is still behind R in this regard, Python offers a wide range of options for creating visualizations.
3. Tableau: Tableau is another popular tool for data science visualization. It provides a user-friendly interface and offers a variety of interactive and dynamic visualizations for exploring and presenting data.

c) What are the important aspects of Data Visualization in the field of Data Science?
Ans.

Aspects of Data Visualization in Data Science

Data visualization is an important aspect of data science that helps in understanding and communicating complex information through visual representations. It involves creating visualizations such as charts, graphs, and maps to explore patterns, trends, and relationships in data. Effective data visualization enhances data analysis, aids in decision-making, and facilitates the communication of insights to stakeholders.

Some important aspects of data visualization in the field of data science include:

1. Data Exploration: Data visualization allows data scientists to explore and understand the characteristics of the data. It helps in identifying outliers, patterns, and trends that may not be apparent in raw data. Visualizations provide a visual summary of the data, making it easier to identify key insights and relationships.
2. Communication of Insights: Data visualization helps in effectively communicating insights and findings to stakeholders. Visual representations make complex information more accessible and understandable to a wider audience. It enables data scientists to present their analysis and conclusions in a clear and concise manner, facilitating decision-making and action.
3. Interactive Visualizations: Interactive visualizations allow users to interact with the data, enabling them to explore different aspects and drill down into specific details. Interactive features such as filtering, zooming, and sorting enhance the user experience and enable deeper analysis of the data.
4. Storytelling: Data visualization can be used to tell a story with data. By combining visual elements, narrative, and context, data scientists can create compelling visualizations that engage and captivate the audience. Storytelling through data visualization helps in conveying the message and insights in a memorable and impactful way.

d) Why Statistical Power is Needed?
ANS.

Statistical power is a measure of the ability of a statistical test to detect an effect or relationship if it exists in the population. It is important in data analysis and hypothesis testing as it determines the likelihood of correctly rejecting the null hypothesis when it is false.

Statistical power is needed for several reasons:

1. Detecting True Effects: Statistical power ensures that a study has a high probability of detecting true effects or relationships in the data. It helps in avoiding false negative results, where a significant effect is missed due to insufficient power.
2. Reducing Type II Errors: Type II errors occur when a study fails to reject the null hypothesis when it is false. By increasing statistical power, the likelihood of committing a Type II error is reduced, increasing the confidence in the results.
3. Sample Size Determination: Statistical power is used in determining the sample size required for a study. It helps in estimating the number of participants or observations needed to achieve a desired level of power, ensuring that the study is adequately powered to detect meaningful effects.
4. Generalizability of Results: Adequate statistical power increases the generalizability of study findings to the larger population. It ensures that the results are not limited to the specific sample studied and can be applied to a broader context.

In summary, statistical power is needed in data analysis to ensure the accuracy and reliability of results, reduce errors, determine sample size, and enhance the generalizability of findings.

Statistical Power: Why is it Needed?

Statistical power is a crucial concept in statistical analysis. It measures the ability of a statistical test to detect an effect or relationship if it truly exists in the population. In other words, it determines the probability of correctly rejecting a null hypothesis when it is false.

Having sufficient statistical power is important because it helps researchers avoid Type II errors, which occur when a true effect or relationship is not detected. By ensuring adequate power, researchers can increase the chances of finding meaningful results and drawing accurate conclusions from their data.

Statistical power is influenced by several factors, including sample size, effect size, and significance level. Increasing the sample size and effect size, or decreasing the significance level, can enhance the power of a statistical test.

In summary, statistical power is needed to ensure that research studies have the ability to detect true effects or relationships. It helps researchers make reliable inferences and draw valid conclusions from their data

e) What is the difference between probability sampling and non-probability sampling?
ANS.

Difference between Probability Sampling and Non-Probability Sampling

Probability Sampling:

- In probability sampling, each element of the population has a known and non-zero probability of being selected for the sample.
- This method is usually preferred because its properties, such as bias and sampling error, are usually known.
- It ensures that the sample is representative of the population as a whole.

Non-Probability Sampling:

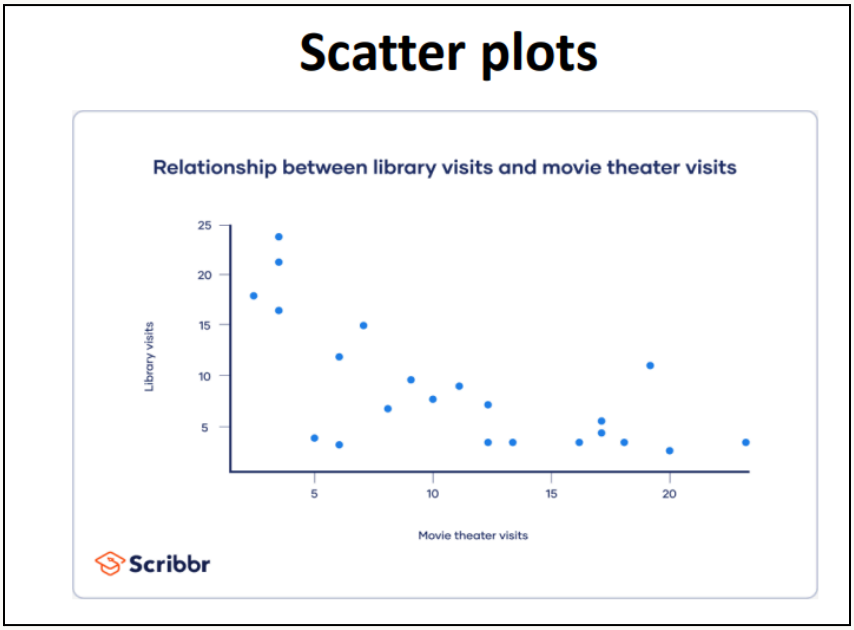
- In non-probability sampling, some elements of the population may not be selected, and there is a risk of the sample being non-representative of the population.
- Non-probability sampling is based on the judgment of the analyst and does not assign known probabilities to each element.
- It can be used when probability sampling is not possible or when it is more cost-effective to use non-random methods.

Overall, the main difference between probability sampling and non-probability sampling lies in the assignment of known probabilities to each element of the population. Probability sampling ensures representativeness, while non-probability sampling relies on the judgment of the analyst and may introduce biases.

Q: 2 Solve the following.
a) Explain any four Data Plotting Types and their Significance in data visualization.[8] b) 9 people take a test. Their scores out of 100 are:
56,79,77,48,90,68,79,92,71
Work out the mean, median, and mode of their scores. [7]
ANS.

Data Plotting Types and their Significance in data visualization

1. Scatter Plot: A scatter plot is used to visualize the relationship between two variables. It helps in identifying patterns, trends, and correlations between the variables. By plotting the data points on a chart, we can determine if there is a positive, negative, or no relationship between the variables.



2. Bar Chart: A bar chart is used to compare categorical data. It represents the data using rectangular bars, where the length of each bar corresponds to the frequency or proportion of the category. Bar charts are effective in displaying comparisons and identifying the highest or lowest values among categories.
3. Line Graph: A line graph is used to show the trend or change in data over time. It is particularly useful for tracking continuous data and visualizing patterns or trends. By connecting data points with lines, we can observe the overall direction and magnitude of change.
4. Histogram: A histogram is used to display the distribution of continuous data. It divides the data into intervals or bins and represents the frequency or proportion of data points falling within each bin. Histograms help in understanding the shape, central tendency, and spread of the data.

Mean, Median, and Mode of Test Scores

Given scores: 56, 79, 77, 48, 90, 68, 79, 92, 71

Mean: To calculate the mean, we sum up all the scores and divide by the total number of scores. Sum of scores = 56 + 79 + 77 + 48 + 90 + 68 + 79 + 92 + 71 = 660. Total number of scores = 9. Mean = 660 / 9 = 73.33

Median: To find the median, we arrange the scores in ascending order and find the middle value. Arranged scores: 48, 56, 68, 71, 77, 79, 79, 90, 92. Median = 77

Mode: The mode is the score that appears most frequently. In this case, the mode is 79 as it appears twice, which is more than any other score.

Mean: 73.33 Median: 77 Mode: 79

Q: 3 Solve the following.
a) What are the Advantages and disadvantages of non probability sampling.[8]
ANS.
Advantages of Non-Probability Sampling:

1. Cost-effective: Non-probability sampling methods, such as volunteer sampling, can be cheaper to implement compared to probability sampling. This makes it a viable option when budget constraints are a concern.
2. Easy and quick: Non-probability sampling methods are relatively easy to implement and can provide quick results. For example, posting a survey form on a Facebook group can quickly gather responses from willing participants.
3. Initial validation: Non-probability sampling can serve as an initial validation step to gauge interest before investing in more expensive sampling methods. It can help determine if there is sufficient interest to pursue further research.

Disadvantages of Non-Probability Sampling:

1. Non-representative sample: Non-probability sampling carries a risk of producing a sample that is not representative of the entire population. This can introduce bias and limit the generalizability of the findings.
2. Limited control over sample selection: With non-probability sampling, researchers have limited control over who participates in the study. This can result in an oversampling of certain groups or individuals, leading to skewed results.
3. Sampling error: Non-probability sampling methods may introduce a sampling error, making it difficult to accurately estimate population characteristics. This can affect the reliability and validity of the findings.
4. Lack of generalizability: Due to the non-random nature of non-probability sampling, the findings may not be generalizable to the larger population. This can limit the applicability of the results beyond the specific sample studied.
5. Potential for bias: Non-probability sampling methods can be susceptible to various biases, such as self-selection bias or response bias. These biases can distort the findings and compromise the integrity of the research.

In summary, non-probability sampling methods offer cost and time efficiency, as well as an initial validation step. However, they come with the risk of producing non-representative samples, limited control over sample selection, potential sampling errors, lack of generalizability, and biases. Researchers should carefully consider these advantages and disadvantages when choosing the appropriate sampling method for their study.

b) Define: i) Priori Power Analysis ii)z-score iii)Data import iv) post-hoc Power Analysis v) Population vi) Sample vii) Sampling [7]
ANS.

i) Priori Power Analysis A priori power analysis is a statistical method used to determine the sample size needed for a study before data collection begins. It helps researchers estimate the statistical power of their study, which is the probability of detecting a true effect if it exists. By conducting a priori power analysis, researchers can ensure that their study has enough participants to detect meaningful results.

ii) z-score A z-score, also known as a standard score, is a statistical measure that indicates how many standard deviations an individual data point is from the mean of a distribution. It is calculated by subtracting the mean from the data point and dividing the result by the standard deviation. Z-scores are commonly used in hypothesis testing and determining the relative position of a data point within a distribution.

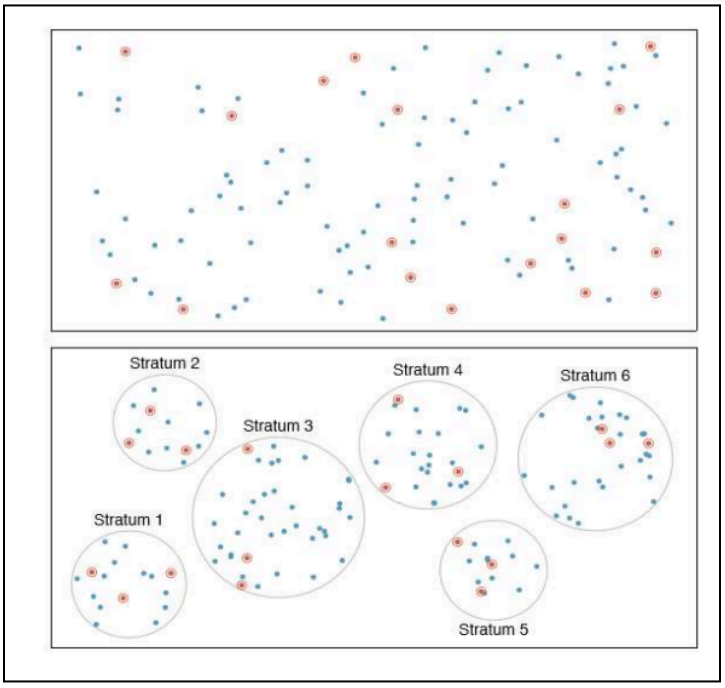
iii) Data import Data import refers to the process of bringing data from an external source into a software or program for analysis. It involves transferring data from a file or database into a format that can be easily manipulated and analyzed. Data import is an essential step in data analysis as it allows researchers to work with the relevant data in their chosen software or program.

iv) post-hoc Power Analysis Post-hoc power analysis is a statistical method used to determine the statistical power of a study after data collection and analysis have been completed. It is conducted to assess whether the sample size used in the study was sufficient to detect meaningful effects. Post-hoc power analysis can help researchers understand the limitations of their study and the reliability of their findings.

v) Population In statistics, a population refers to the entire group of individuals, objects, or events that a researcher is interested in studying. It is the larger group from which a sample is drawn. The characteristics and parameters of a population are of interest to researchers, but it is often impractical or impossible to collect data from the entire population. Therefore, researchers use sampling techniques to study a representative subset of the population.

vi) Sample A sample is a subset of individuals, objects, or events selected from a larger population for the purpose of data analysis. It is used to make inferences and draw conclusions about the population from which it was drawn. The sample should be representative of the population to ensure the validity of the findings. Sampling methods, such as random sampling or stratified sampling, are used to select the sample from the population.

vii) Sampling Sampling is the process of selecting a subset of individuals, objects, or events from a larger population for the purpose of data analysis. It involves choosing a representative sample that accurately reflects the characteristics of the population. Different sampling methods, such as simple random sampling or cluster sampling, can be used depending on the research objectives and the nature of the population. Sampling is essential in statistical analysis as it allows researchers to make inferences about the population based on the characteristics of the sample.



stratified sampling diagram

Q: 4 Solve the following.
a) Explain linear regression with example.[8]
ANS.

Linear Regression

Linear regression is a statistical technique used to model the relationship between two variables. It assumes a linear relationship between the independent variable (x) and the dependent variable (y). The goal of linear

regression is to find the best-fitting line that minimizes the sum of the squared differences between the observed and predicted values.

For example, let's say we want to predict a person's weight (dependent variable) based on their height (independent variable). We collect data on the heights and weights of several individuals and plot them on a scatter plot. By fitting a straight line to the data points, we can estimate the weight of a person based on their height.

Linear regression is a simple yet powerful tool for predicting continuous outcomes and understanding the relationship between variables. It is widely used in various fields such as economics, finance, and social sciences.

b) What are the benefits and challenges of power analysis? [7]
ANS.

Benefits of Power Analysis

Power analysis can provide several benefits in the field of data analysis. Firstly, it allows researchers to determine the sample size needed to detect a statistically significant effect. This helps in optimizing resources and ensuring that the study has enough power to detect meaningful results. Secondly, power analysis helps in designing experiments by estimating the effect size and variability, which aids in making informed decisions about the study design. Lastly, power analysis can also be used to evaluate the statistical power of existing studies, providing insights into the reliability of their findings.

Challenges of Power Analysis

While power analysis is a valuable tool, it also comes with certain challenges. One of the main challenges is the need for accurate estimation of effect size and variability, which can be difficult in some research domains. Additionally, power analysis assumes certain statistical assumptions, such as normality and independence, which may not always hold true in real-world scenarios. Another challenge is the trade-off between sample size and resources, as increasing the sample size may require more time, effort, and cost. Finally, power analysis relies on the correct specification of the statistical test and hypothesis, and any deviations from these assumptions can affect the accuracy of the analysis.

Q: 5 Solve the following.[15]
a) Why does reproducibility matter? How to make Your Projects Reproducible? Explain. [8]
ANS.

Why does reproducibility matter?

Reproducibility matters for several reasons. Firstly, it allows others to understand and reproduce a researcher's work, increasing the strength of evidence for a phenomenon. Secondly, it promotes compatibility and consistency among similar studies, enabling them to provide evidence in support of or in opposition to a concept. Additionally, reproducibility enhances the utility of studies for meta-analyses, which are important for generalizing and contextualizing findings.

How to make Your Projects Reproducible?

To make your projects reproducible, there are a few key steps to follow. Firstly, ensure that your research is well-documented, with clear explanations of how and why specific analyses were performed. This helps

collaborators, supervisors, and reviewers understand your work. Secondly, share your data and code with others, allowing them to access and reproduce your analyses. This not only facilitates learning and collaboration but also enables quick reconfiguration of previously conducted research tasks for future projects. Finally, strive for rigor in your analyses, following best practices and validating your methods to ensure the reliability and reproducibility of your results.

b) What do you understand as Frequentism? Justify your answer. [7]
ANS.

Frequentism

Frequentism is an approach to statistics that focuses on the frequency or probability of an event occurring. It is based on the idea that repeated observations of an event can provide reliable information about its true probability. In frequentism, probabilities are interpreted as long-run frequencies, meaning that the probability of an event is the proportion of times it would occur in an infinite number of repetitions. This approach does not take into account prior beliefs or subjective probabilities, but relies solely on observed data. Frequentism is often contrasted with Bayesianism, which incorporates prior beliefs and subjective probabilities into statistical inference.

Q: 6 Solve the following.
a) Explain Bayesian Inference.[8]
ANS.

Bayesian Inference

Bayesian Inference is a statistical method used to update our beliefs or knowledge about a hypothesis or event based on new evidence or data. It involves using Bayes' theorem to calculate the probability of a hypothesis being true given the observed data. This method allows us to incorporate prior knowledge or beliefs into our analysis and update them as new data becomes available. Bayesian Inference is widely used in various fields, including data science, machine learning, and decision-making processes.

b) Explain Data Simulation.[7]
ANS.

Data Simulation

Data simulation is the process of using a large amount of data to create a virtual representation of real-world conditions. It is used to predict future instances, determine the best course of action, or validate a model. Simulations can be used to verify analytical solutions, experiment with policies before implementing them physically, and understand the relationship between different variables in a system. By modifying input parameters and examining the results, simulations can provide insights into the behavior of complex systems. Simulated data can be used to produce realistic models, visualize trends and results, and compare different scenarios to make informed decisions. The features of data simulation tools may include a graphical user interface to make it accessible to a wider audience and facilitate the formulation and execution of simulations.

Q: 7 Write a short note on: [15]
a) simple Random Sampling [5]
ANS.

Simple Random Sampling

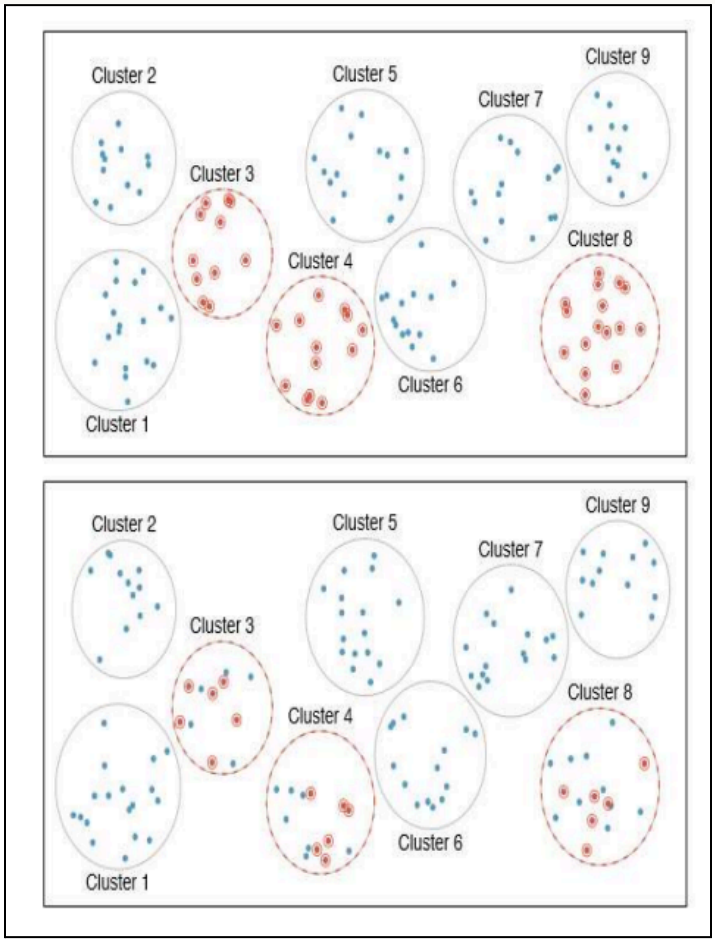
Simple random sampling is a method of selecting a sample from a population where each case has an equal chance of being included in the sample. It is considered the most intuitive form of random sampling. For example, in the context of Major League Baseball (MLB) players, a simple random sample could be taken by writing the names of all players onto slips of paper, mixing them up in a bucket, and drawing out slips until the desired sample size is reached. This ensures that each player has an equal chance of being selected for the sample.

b) Multi-stage sampling [5]
ANS.

Multi-stage sampling

Multi-stage sampling is a more complex form of cluster sampling. It involves dividing the larger population into clusters and then breaking out second-stage clusters based on a secondary factor. These clusters are then sampled and analyzed. This process can continue with multiple subsets being identified, clustered, and analyzed.

In multi-stage sampling, the population is divided into clusters, and then subsets of each cluster are randomly selected to be included in the sample. This method allows for a more detailed analysis of the population by analyzing different subsets within the clusters.



c) Quota sampling [5]
ANS.

Quota Sampling

Quota sampling is a non-probability sampling method where the researcher selects participants based on specific characteristics or quotas. The population is divided into groups or strata, and a predetermined number of participants is selected from each group to meet the desired quotas. This method allows for control over the composition of the sample, but it may introduce bias if the quotas are not representative of the population. Quota sampling is often used when probability sampling methods are not feasible or practical.

PYQ 8-12-2022

Q: 1 Solve any three.
a) What are the different data structures in R? Briefly explain about them.
ANS.

Data Structures in R

R programming language provides several data structures that are used to store and manipulate data. These data structures include vectors, matrices, arrays, lists, and data frames.

1. Vectors: Vectors are one-dimensional arrays that can hold elements of the same data type. They can be created using the
   `c()`
   function and can store numeric, character, or logical values.
2. Matrices: Matrices are two-dimensional arrays that can store elements of the same data type. They can be created using the
   `matrix()`
   function and are useful for performing matrix operations.
3. Arrays: Arrays are multi-dimensional data structures that can store elements of the same data type. They can be created using the
   `array()`
   function and are useful for storing and manipulating data in more than two dimensions.
4. Lists: Lists are a collection of objects that can be of different data types. They can be created using the
   `list()`
   function and are useful for storing heterogeneous data.
5. Data Frames: Data frames are two-dimensional data structures that can store data of different data types. They can be created using the
   `data.frame()`
   function and are commonly used for data analysis and manipulation.

These data structures in R provide flexibility and efficiency in handling and analyzing data, making R a powerful language for statistical computing and data analysis.

b) Give any Top 3 Data Science Visualization Tools.
ANS.

Top 3 Data Science Visualization Tools

1. R: R is a powerful tool for data visualization with packages like ggplot2, ggvis, and lattice. These packages make data visualization easier and provide a wide range of options for creating visually appealing plots and charts.

2. Python: Python is catching up with data visualization tools like Bokeh and Matplotlib. While it may still be behind R in this regard, Python offers a growing number of packages that allow for creating interactive and visually appealing visualizations.
3. Tableau: Tableau is a popular data visualization tool that provides a user-friendly interface for creating interactive and dynamic visualizations. It offers a wide range of features and options for exploring and presenting data in a visually appealing manner.

c) 9 people take a test. Their scores out of 100 are:
56,79,77,48,90,68,79,92,71
Work out the mean, median, and mode of their scores.
ANS.

Mean: To calculate the mean, we add up all the scores and divide the sum by the total number of scores. In this case, the sum of the scores is 56 + 79 + 77 + 48 + 90 + 68 + 79 + 92 + 71 = 660. Since there are 9 scores, the mean is 660/9 = 73.33.

Median: The median is the middle value when the scores are arranged in ascending order. In this case, when we arrange the scores in ascending order, we get 48, 56, 68, 71, 77, 79, 79, 90, 92. Since there are 9 scores, the middle value is the 5th score, which is 77.

Mode: The mode is the score that appears most frequently. In this case, the score that appears most frequently is 79, so the mode is 79.

d) Define: i. Population ii. Sample iii. Sampling iv.Null Hypothesis v. Alternate hypothesis
ANS.
Population: The population refers to the entire group of individuals, objects, or events that we are interested in studying or making inferences about. It is the larger group from which a sample is drawn.

Sample: A sample is a subset of the population that is selected for study or analysis. It is a smaller representative group that is used to make inferences or draw conclusions about the larger population.

Sampling: Sampling is the process of selecting a subset of individuals, objects, or events from the population to be included in the sample. It involves choosing a representative group that can provide insights into the characteristics or behavior of the entire population.

Null Hypothesis: The null hypothesis is a statement or assumption that suggests there is no significant difference or relationship between variables in a study. It is often used in hypothesis testing to determine if there is enough evidence to reject the null hypothesis in favor of an alternative hypothesis.

Alternative Hypothesis: The alternative hypothesis is a statement or assumption that suggests there is a significant difference or relationship between variables in a study. It is the opposite of the null hypothesis and is used to support a researcher's hypothesis or claim.

e) How Data Import works?
ANS.

Data Import in Computational Data Analytics

Data import is an essential step in computational data analytics. Models require that data sets be imported to the model for analysis and processing. The process of data import involves bringing in external data into the computational environment, allowing it to be used for simulations, modeling, and analysis.

In the context of data simulation, data import enables the incorporation of external data sets into the simulation model. This allows for the creation of more realistic and accurate simulations by using real-world data. The imported data can be used to generate insights, test hypotheses, and make predictions based on the simulated scenarios.

The tools used for data simulation often provide features for data import and export. These features allow users to easily import data sets into the simulation model and export the results of the simulation for further analysis or visualization. The data import functionality is designed to be user-friendly, enabling users with varying levels of expertise to import and utilize data in their simulations.

Overall, data import plays a crucial role in computational data analytics, enabling the integration of external data sets into simulation models for analysis and decision-making.

Q: 2 Solve the following.
a) Explain four different types of non-probability sampling. [8]
ANS.

Four Types of Non-Probability Sampling

1. Judgement Sampling: This method involves selecting a sample based on existing domain knowledge. For example, if you want to survey potential customers for a new coding online course, you might choose individuals based on your understanding of the type of people who would be interested.
2. Volunteer Sampling: This widely used method involves collecting data from individuals who voluntarily participate, such as when you post a survey form on a Facebook group and ask people to fill it. However, this method can introduce bias as it may oversample individuals who are on Facebook, like you, or have enough free time to fill out the form.
3. Convenience Sampling: This type of sampling involves selecting individuals who are readily available and convenient to access. For instance, if you conduct a survey at a shopping mall, you are sampling individuals who happen to be present at that location.
4. Snowball Sampling: In this method, initial participants are selected, and then they refer or recruit additional participants. This approach is often used when it is difficult to identify and access the target population directly.

These four types of non-probability sampling methods provide different ways to gather data, but it's important to note that they may introduce biases and limitations compared to probability sampling methods.

b) What is model selection and explain two different techniques of model selection.[7]
ANS.

Model Selection

Model selection is the process of choosing the most appropriate model from a set of candidate models for a given problem. It involves evaluating different models based on their performance and selecting the one that best fits the data and provides accurate predictions.

Techniques of Model Selection

1. Cross-validation: Cross-validation is a technique used to assess the performance of a model on an independent dataset. It involves dividing the available data into multiple subsets, training the model on a subset, and evaluating its performance on the remaining subset. This process is repeated multiple times, and the average performance is used to select the best model.
2. Information criteria: Information criteria, such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), are statistical measures used to compare different models based on their goodness of fit and complexity. These criteria penalize models with higher complexity, encouraging the selection of simpler models that still provide good fit to the data.

These techniques help in selecting the most suitable model by considering factors such as model performance, complexity, and generalizability.

Q: 3 Solve the following.
a) Explain any four Data Plotting Types and their Significance in data visualization.[8]
ANS.

Data Plotting Types and their Significance in Data Visualization

1. Scatter Plot: A scatter plot is a chart that displays the relationship between two or three variables. It helps visualize the strength of the relationship between the variables by plotting one variable along the x-axis and another along the y-axis. Scatter plots are useful for identifying patterns, trends, and correlations in the data.
2. Line Plot: A line plot is a chart that represents data points connected by straight lines. It is commonly used to show the trend or change in a variable over time. Line plots are effective in visualizing the progression or fluctuation of data and can be used to identify patterns, trends, or seasonality.
3. Bar Plot: A bar plot, also known as a bar chart, uses rectangular bars to represent data. It is commonly used to compare and display categorical data. Each bar represents a category, and the height of the bar corresponds to the value or frequency of that category. Bar plots are useful for comparing different categories or groups and identifying the highest or lowest values.
4. Histogram: A histogram is a graphical representation of the distribution of numerical data. It divides the data into intervals or bins and displays the frequency or count of data points within each bin. Histograms are useful for understanding the shape, central tendency, and variability of the data distribution.

These data plotting types are significant in data visualization as they provide visual representations that help in understanding and interpreting data patterns, trends, and relationships. They enable analysts to identify outliers, detect patterns, compare categories, and gain insights from the data. By using these plotting types, data can be presented in a more intuitive and understandable manner, facilitating effective decision-making and communication.

b) Write a short note on power analysis.[7]
ANS.

Power Analysis

Power analysis is a statistical method used to determine the sample size needed for a study to detect a significant effect. It helps researchers determine the probability of correctly rejecting the null hypothesis when it is false. By conducting a power analysis, researchers can ensure that their study has enough statistical power to detect meaningful effects and avoid wasting resources on underpowered studies. Power analysis takes into account factors such as effect size, significance level, and desired power level to provide guidance on sample size determination.

Q: 4 Solve the following.
a) Explain linear regression with example.[8]
ANS.

Linear Regression

Linear regression is a statistical technique used to model the relationship between two variables. It assumes a linear relationship between the independent variable (x) and the dependent variable (y). The goal of linear regression is to find the best-fitting line that minimizes the sum of the squared differences between the observed and predicted values.

For example, let's say we want to predict a person's weight (dependent variable) based on their height (independent variable). We collect data on the heights and weights of several individuals and plot them on a scatter plot. By fitting a straight line to the data points, we can estimate the weight of a person based on their height.

Linear regression is a simple yet powerful tool for predicting continuous outcomes and understanding the relationship between variables. It is widely used in various fields such as economics, finance, and social sciences.

b) Explain Data Analytics Life cycle.[7]
ANS.

Data Analytics Life Cycle

The data analytics life cycle consists of several stages that are followed to extract meaningful insights from data. These stages include data collection, data preparation, data analysis, and data visualization.

1. Data Collection: In this stage, relevant data is gathered from various sources such as databases, spreadsheets, or online platforms. The data collected should be accurate, complete, and representative of the problem at hand.
2. Data Preparation: Once the data is collected, it needs to be cleaned and transformed into a suitable format for analysis. This involves removing any inconsistencies, missing values, or outliers that may affect the accuracy of the results.
3. Data Analysis: After the data is prepared, various statistical techniques and algorithms are applied to uncover patterns, trends, and relationships within the data. This analysis helps in drawing meaningful conclusions and making data-driven decisions.
4. Data Visualization: The insights obtained from the data analysis are then visualized using charts, graphs, or other visual representations. This helps in presenting the findings in a clear and understandable manner, making it easier for stakeholders to interpret and act upon the results.

By following this data analytics life cycle, organizations can effectively leverage their data to gain valuable insights, improve decision-making processes, and drive business growth.

Q: 5 Solve the following.
a) Explain Bayesian Inference.[8]
ANS.

Bayesian Inference

Bayesian Inference is a statistical method used to update our beliefs or knowledge about a hypothesis or parameter based on new evidence or data. It involves using Bayes' theorem to calculate the posterior probability of a hypothesis given the observed data. This method allows us to incorporate prior knowledge or beliefs into our analysis and update them as we gather more information. Bayesian Inference is widely used in various fields, including data science, machine learning, and decision-making processes.

b) "Despite the wide-reaching benefits that come with using big data analytics, its use also comes with challenges" Explain.[7]
ANS.

Challenges of Using Big Data Analytics

Big data analytics offers numerous benefits, but it also presents certain challenges. These challenges include:

1. Data Volume: Big data analytics deals with large volumes of data, which can be difficult to manage and analyze efficiently. Processing and storing such massive amounts of data require specialized systems and software.
2. Data Sampling: Sampling is often used in big data analytics to analyze representative subsets of the data. However, selecting an appropriate sample size and avoiding sampling errors can be challenging.
3. Data Manipulation and Interpretation: As the size of the data sample increases, manipulating and interpreting the data becomes more complex. It may require additional computational resources and expertise to handle and extract meaningful insights from large datasets.
4. Data Quality: Ensuring the quality and accuracy of the data used for analysis is crucial. Big data analytics relies on the assumption that the data is reliable and representative. However, data quality issues, such as incomplete or inconsistent data, can affect the accuracy of the analysis results.
5. Privacy and Security: Big data analytics involves handling sensitive and personal information. Ensuring data privacy and security is a significant challenge, as unauthorized access or data breaches can have severe consequences.

Despite these challenges, the benefits of using big data analytics, such as identifying trends, making predictions, and solving complex problems, make it a valuable tool for organizations. By addressing these challenges effectively, organizations can harness the power of big data analytics to gain valuable insights and make informed decisions.

Q: 6 Solve the following.
a) The values of y and their corresponding values of y are shown in the table below

X	0	1	2	3	4
y	2	3	5	4	6

i. Find the least square regression line $y = ax + b$.
ii.Estimate the value of y when $x = 10$.
ANS.

i. Find the least square regression line $y = ax + b$.

To find the least square regression line, we need to calculate the values of a and b. The formula for calculating a is given by:

$a = (n\Sigma xy - \Sigma x\Sigma y) / (n\Sigma x^2 - (\Sigma x)^2)$

Using the given values of x and y, we can calculate the necessary sums:

Σx = 0 + 1 + 2 + 3 + 4 = 10 Σy = 2 + 3 + 5 + 4 + 6 = 20 Σxy = (02) + (13) + (25) + (34) + (4*6) = 40 Σx^2 = (0^2) + (1^2) + (2^2) + (3^2) + (4^2) = 30

Substituting these values into the formula, we can calculate a:

a = (540 - 1020) / (5*30 - 10^2) = 10/5 = 2

Now, we can calculate b using the formula:

b = (Σy - aΣx) / n

Substituting the values, we get:

b = (20 - 2*10) / 5 = 0

Therefore, the least square regression line is y = 2x.

ii. Estimate the value of y when x = 10.

Since the given data only includes values of x up to 4, we cannot directly estimate the value of y when x = 10 using the least square regression line. We would need additional data points beyond x = 4 to make a valid estimation.

b) What are the four types of data analysis?Explain them. [7]
ANS.

Four Types of Data Analysis

1. Descriptive Analysis: Descriptive analysis involves summarizing and describing the main characteristics of a dataset. It focuses on understanding the distribution, central tendency, and variability of the data. This type of analysis helps in gaining insights into the dataset and identifying patterns or trends.
2. Diagnostic Analysis: Diagnostic analysis aims to understand the reasons behind certain patterns or trends observed in the data. It involves investigating relationships between variables and identifying the factors that contribute to specific outcomes. Diagnostic analysis helps in identifying the root causes of problems or issues.
3. Predictive Analysis: Predictive analysis uses historical data to make predictions or forecasts about future events or outcomes. It involves building statistical models or machine learning algorithms to analyze patterns in the data and make predictions. Predictive analysis helps in making informed decisions and planning for the future.
4. Prescriptive Analysis: Prescriptive analysis goes beyond predicting future outcomes and provides recommendations or actions to optimize results. It combines historical data, predictive models, and optimization techniques to suggest the best course of action. Prescriptive analysis helps in making data-driven decisions and improving overall performance.

These four types of data analysis provide a comprehensive approach to understanding, interpreting, and utilizing data for decision-making and problem-solving purposes.

Q: 7 Write a short note on: [15]
a) Z Test.[5]
ANS.

Z Test

The given document does not provide any information about the Z test. I'm sorry, but I don't know the answer to your question from the document content provided.

b) simple Random Sampling[5]
ANS.

Simple Random Sampling

Simple random sampling is a method of selecting a sample from a population where each case has an equal chance of being included in the sample. It is considered the most intuitive form of random sampling. For example, in the context of Major League Baseball (MLB) players, a simple random sample could be taken by writing the names of all players onto slips of paper, mixing them up in a bucket, and drawing out slips until the desired sample size is reached. This ensures that each player has an equal probability of being included in the sample.

c) Central Tendency measures[5]
ANS.

Measures of Central Tendency

Measures of central tendency estimate the center or average of a data set. The mean, median, and mode are three ways of finding the average. The mean, or M, is the most commonly used method for finding the average. To find the mean, simply add up all response values and divide the sum by the total number of responses. The total number of responses or observations is called N.

Mean

The mean is calculated by dividing the sum of all values by the total number of responses. In the given data set of 15, 3, 12, 0, 24, and 3, the sum of all values is 57 and the total number of responses is 6. Therefore, the mean is 9.5.

Median

The median is the middle value in a data set when the values are arranged in ascending or descending order. In the given data set, the median is 7.5.

Mode

The mode is the value that appears most frequently in a data set. In the given data set, the mode is 3.

These measures of central tendency provide information about the average or typical value of a data set. The mean is influenced by outliers, while the median and mode are more resistant to outliers.

UT2 2024
Note: Solve any Five questions each carry 5 marks.
1. What is diagnostic analytics? Explain the purpose of diagnostic analytics in detailed
ANS.

Diagnostic Analytics

Diagnostic analytics is a branch of data analytics that focuses on examining data sets to identify the causes of past events or outcomes. It involves analyzing historical data to understand why certain events occurred and what factors contributed to them. The purpose of diagnostic analytics is to gain insights into the root causes of problems or successes, enabling organizations to make informed decisions and take appropriate actions to improve future outcomes. It helps in identifying patterns, trends, and relationships in data to provide a deeper understanding of the factors that drive specific outcomes. By analyzing data from various sources, diagnostic analytics helps in uncovering hidden insights and providing valuable information for decision-making.

2. Explain the working of diagnostic analytics in detailed.
ANS.

Working of Diagnostic Analytics

Diagnostic analytics is a branch of data analytics that focuses on understanding the reasons behind certain outcomes or events. It involves analyzing historical data to identify patterns, correlations, and causal relationships. By examining the data, diagnostic analytics helps in determining why certain events occurred and what factors contributed to them.

In the context of computational data analytics, diagnostic analytics utilizes specialized systems and software to analyze large data sets. It applies algorithms and statistical techniques to uncover insights and explanations for specific outcomes. By examining the data in detail, diagnostic analytics helps in identifying the root causes of problems or successes.

The process of diagnostic analytics involves several steps. First, the data is collected and organized into a structured format. Then, various statistical techniques and algorithms are applied to analyze the data. These techniques may include regression analysis, hypothesis testing, and data visualization.

Through diagnostic analytics, organizations can gain a deeper understanding of their data and make informed decisions. It helps in identifying trends, anomalies, and outliers that may have contributed to specific outcomes. By understanding the underlying factors, organizations can take corrective actions or replicate successful strategies.

Overall, diagnostic analytics plays a crucial role in uncovering the "why" behind data patterns and outcomes. It helps organizations gain insights into their data and make data-driven decisions for improved performance and problem-solving.

3. What is probability density estimation? Explain with suitable example.
ANS.

Probability Density Estimation

Probability density estimation is a statistical technique used to estimate the probability density function (PDF) of a random variable based on a given set of data. The PDF represents the likelihood of a random variable taking on a specific value.

One common method of probability density estimation is the kernel density estimation (KDE) technique. In KDE, a kernel function is used to estimate the PDF by placing a kernel at each data point and summing them up to create a smooth estimate of the underlying distribution.

For example, let's say we have a dataset of heights of individuals. We can use probability density estimation to estimate the PDF of the height variable. By applying the KDE technique, we can obtain a smooth curve that represents the likelihood of observing different heights in the population. This can be useful for various applications, such as understanding the distribution of heights in a population or making predictions based on the estimated PDF.

4. How Likelihood frequentist and the Machine Learning are ascociated with each other.
ANS.

Likelihood and Frequentist Approach

The likelihood and frequentist approach are both statistical methods used in data analysis. The likelihood approach focuses on estimating the parameters of a statistical model based on the observed data. It calculates the probability of observing the data given a specific set of parameter values. On the other hand, the frequentist approach treats the parameters as fixed and unknown, and aims to estimate them based on the frequency of occurrence of the observed data. Both approaches are used in machine learning to make predictions and infer insights from data.

5. What are the assumptions of linear regression how will you deal with the violation of normality, multicolonearity assumptions.
ANS.

Assumptions of Linear Regression:

1. Linearity: The relationship between the independent variables and the dependent variable is linear.
2. Independence: The observations are independent of each other.
3. Homoscedasticity: The variance of the errors is constant across all levels of the independent variables.
4. Normality: The errors are normally distributed.

Dealing with Violation of Normality Assumption:

If the normality assumption is violated, you can consider transforming the data using techniques like logarithmic or square root transformations. Alternatively, you can use robust regression methods that are less sensitive to the normality assumption, such as robust standard errors or quantile regression.

Dealing with Violation of Multicollinearity Assumption:

If the multicollinearity assumption is violated, you can take the following steps:

1. Remove one or more highly correlated independent variables from the model.
2. Combine the highly correlated variables into a single composite variable.
3. Collect more data to reduce the impact of multicollinearity.
4. Use regularization techniques like ridge regression or lasso regression to handle multicollinearity.

Remember that the specific approach to dealing with violations of assumptions may vary depending on the nature and context of the data.

6. Describe the regularization technique of Linear Regression model.
ANS.

Regularization Technique in Linear Regression Model

Regularization is a technique used in linear regression models to prevent overfitting and improve the model's generalization ability. It involves adding a penalty term to the loss function, which helps to control the complexity of the model.

There are two commonly used regularization techniques in linear regression: L1 regularization (Lasso) and L2 regularization (Ridge).

L1 regularization adds the absolute values of the coefficients as the penalty term, which encourages sparsity in the model by shrinking some coefficients to zero. This can be useful for feature selection.

L2 regularization adds the squared values of the coefficients as the penalty term, which encourages small and smooth coefficients. It helps to reduce the impact of outliers and stabilize the model.

Both regularization techniques help to prevent overfitting by adding a penalty for complex models, but they have different effects on the coefficients. L1 regularization tends to produce sparse models with some coefficients set to zero, while L2 regularization tends to shrink all coefficients towards zero. The choice between L1 and L2 regularization depends on the specific problem and the desired characteristics of the model.

7. For the following data, find the linear regression.
6 8 5 10
Ans.

UT1 2024
Note: Solve any Five questions each carry 5 marks.
1. What is Computational data Analytics and Computational Thinking? Explain the steps of Computational Thinking in detailed
ANS.

Computational Data Analytics is a field that allows students to specialize in data science, including machine learning, deep learning, natural language processing, artificial intelligence, visualization, databases, and high-performance computing. It enables scientific discovery by using algorithms to identify patterns and anomalies in data, test hypotheses, create models, and quantify uncertainties.

Computational Thinking is a problem-solving technique that involves four steps: abstraction, automation, analysis, and evaluation.

- Abstraction involves formulating the problem by breaking it down into smaller, more manageable parts.
- Automation is the process of expressing a solution to the problem using algorithms or code.
- Analysis refers to executing the solution and evaluating its effectiveness.
- Evaluation involves assessing the solution's outcomes and making improvements if necessary.

These steps help individuals approach complex problems by decomposing them, recognizing patterns, abstracting essential information, and developing algorithms to solve them. Computational thinking is not limited to computer programming and can be practiced in various ways, such as developing a chess strategy, making and reading maps, or organizing tasks into manageable daily lists. It is a valuable skill for problem-solving in real-world scenarios, from solving math problems to writing a book report.

2. What is R-programming language? Why R is used as a leading tool, explain with its statistical features.

ANS.

Introduction to R Programming Language

R is an open-source programming language widely used as a statistical software and data analysis tool. It comes with a command-line interface and is available on platforms like Windows, Linux, and macOS. R is considered a cutting-edge tool in the field of data analysis.

Why R is Used as a Leading Tool

R programming language is used as a leading tool for machine learning, statistics, and data analysis. It offers the ability to create objects, functions, and packages easily. R is platform-independent, meaning it can be applied to all operating systems. It is also an open-source free language, allowing anyone to install it without purchasing a license. Additionally, R can integrate with other languages like C and C++, enabling easy interaction with various data sources and statistical packages.

Statistical Features of R

R is known for its comprehensive statistical analysis capabilities. It often introduces new technology and concepts first, making it a preferred choice for statistical analysis. Being an open-source language, R can be run anywhere and at any time. It is suitable for GNU/Linux and Windows operating systems and is cross-platform, running on any operating system. Moreover, R has a vibrant community of users who contribute new packages, bug fixes, and code enhancements.

In summary, R programming language is widely used for its statistical features, including its ability to create objects, functions, and packages easily. It is platform-independent, open-source, and can integrate with other languages. R is currently one of the most requested programming languages in the Data Science job market, making it a leading tool in the field of machine learning, statistics, and data analysis.

3. What are measures of central tendency? For the following car data, calculate mean of horsepower of car among the population of car. Also find the centre value of mpg among the population of car. Give justification of both.
(Cars,mpg,cyl,disp,hp,drat)
(A,21,6,160,110,3.9)
(B,21,5,160,110,3.9)
(C,22.8,4,108,93,3.85)
(D,21.3,6,108,96,3)
(E,23,4,150,90,4)
(F,23,6,108,110,3.9)
(G,23,4,160,110,3.9)
(H,23,6,160,110,3.9)
ANS.

Measures of Central Tendency

Measures of central tendency estimate the center or average of a data set. They help us understand the typical or representative value of a variable. The three main measures of central tendency are the mean, median, and mode.

Mean of Horsepower

To calculate the mean of horsepower among the population of cars, we need to add up all the horsepower values and divide the sum by the total number of cars. However, the given data does not include the horsepower values for each car. Therefore, we cannot calculate the mean of horsepower using the given data.

Center Value of MPG

To find the center value of MPG (miles per gallon) among the population of cars, we can use the median. The median is the middle value when the data is arranged in ascending or descending order. In this case, the MPG values are already given in the data set. By arranging the MPG values in ascending order, we find that the center value of MPG is 22.8.

Justification: The median is a robust measure of central tendency that is not affected by outliers. It provides a representative value that is not skewed by extreme values. In this case, the median MPG value of 22.8 represents the typical fuel efficiency of the cars in the population.

4. What is observational sample design, explain in short? Describe the forms of observational studies.
ANS.

Observational Sampling Design

Observational sampling design refers to the method of collecting data in observational studies by monitoring what occurs, without assigning the primary explanatory variable to each subject. Observational studies come in two forms: prospective and retrospective studies.

Prospective Studies

In a prospective study, individuals are identified and followed over a period of time to collect information as events unfold. For example, the Nurses Health Study, started in 1976 and expanded in 1989, is a prospective study that recruits registered nurses and collects data from them using questionnaires. This type of study is useful for assessing the possible influences of behavior on cancer risk.

Retrospective Studies

In a retrospective study, data is collected by looking back at past events or records. This type of study is often used when it is not feasible or ethical to conduct a prospective study. Retrospective studies can provide valuable insights into the relationships between variables, but they are generally only sufficient to show associations and cannot establish causation.

In summary, observational sampling design involves collecting data by monitoring what occurs in observational studies. Prospective studies follow individuals over time, while retrospective studies look back at past events. These studies are useful for assessing associations between variables, but caution must be exercised when making causal conclusions based on observational data.

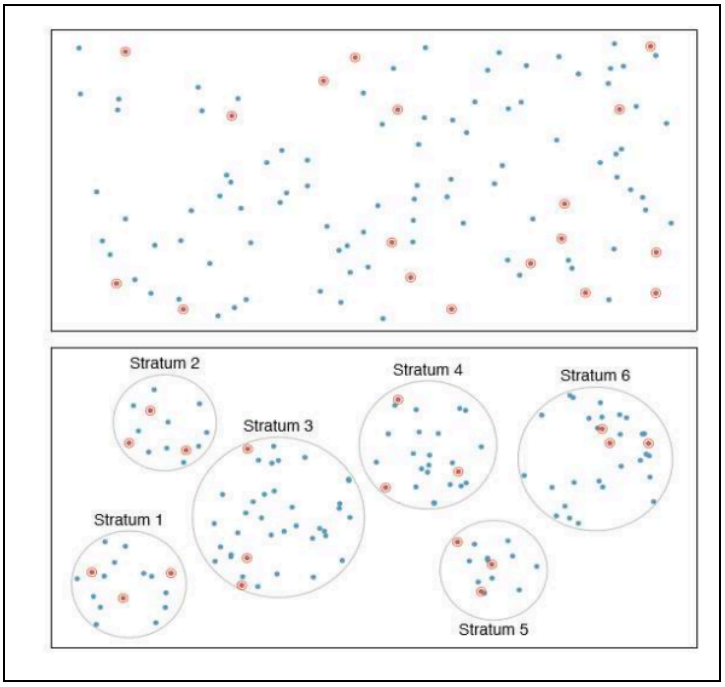
5. What are the three random sampling methods and write a short note on these methods.
ANS.

Three Random Sampling Methods

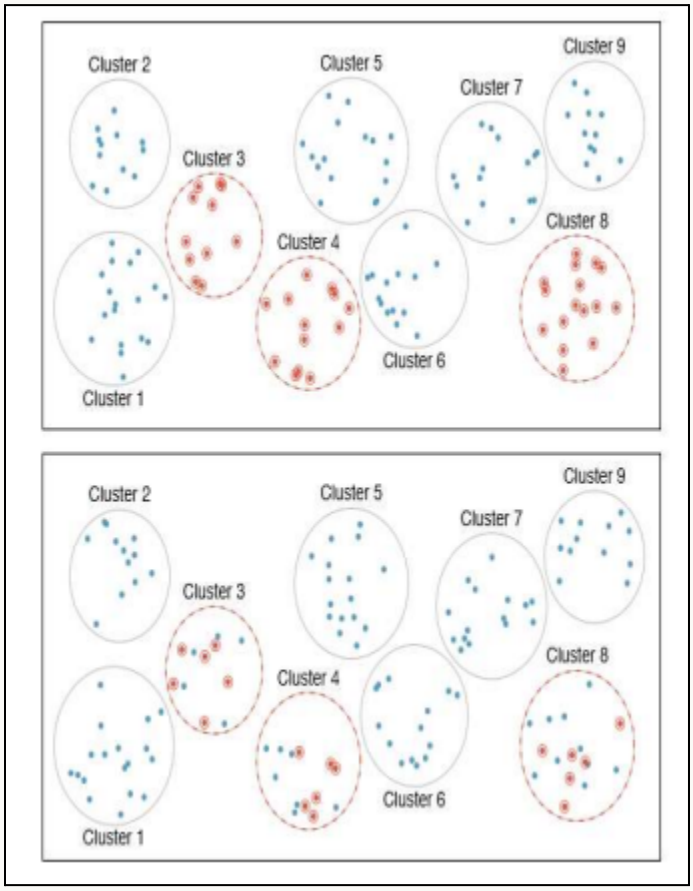
1. Simple Random Sampling: This method involves randomly selecting individuals from a population without any specific criteria. Each individual in the population has an equal chance of being included in the

sample. It is the most intuitive form of random sampling and ensures that each case has an equal probability of being selected.

2. Stratified Sampling: In this method, the population is divided into distinct subgroups or strata based on certain characteristics. Then, a random sample is taken from each stratum in proportion to its size in the population. This ensures representation from each subgroup and allows for more accurate analysis within each stratum.



3. Cluster Sampling: Cluster sampling involves dividing the population into clusters or groups and randomly selecting a few clusters to include in the sample. All individuals within the selected clusters are then included in the sample. This method is useful when it is difficult or impractical to sample individuals directly, such as in large geographical areas or when the population is widely dispersed.



These three random sampling methods provide different approaches to ensure representative samples and reliable statistical analysis. Simple random sampling is straightforward and unbiased, stratified sampling allows for analysis within subgroups, and cluster sampling is useful for large or dispersed populations.

6. What are the barriers to reproducible research? Explain them in short.
ANS.

Barriers to Reproducible Research

1. Complexity: Research often requires specialized knowledge and tools that may not be accessible to everyone. This complexity can hinder reproducibility, but it can be addressed through well-developed tools, protocols, and institutional norms.
2. Technological Change: The hardware and software used for data analysis constantly evolve, making old tools obsolete. This rapid change can make reproducing research challenging. However, researchers can mitigate this barrier by thoroughly annotating code, providing documentation, and using open software when possible.
3. Human Error: Unintentional errors can occur during the data wrangling stage, leading to less reproducible research. To mitigate this, researchers can keep multiple copies of data, document the data conversion process, and double-check a small test set of data before manipulating the entire dataset.
4. Intellectual Property Rights: Concerns over intellectual property rights can also hinder reproducibility. While valid in some scenarios, these concerns can be addressed through appropriate protocols and norms.

Overall, reproducibility can be improved by addressing these barriers through accessible knowledge, well-documented processes, and open practices.

7. Explain the three-step framework for conducting reproducible research.

ANS.

Three-Step Framework for Conducting Reproducible Research

1. Before Data Analysis:
    - This stage involves preparing for data analysis by ensuring that all necessary data and resources are available.
    - Researchers should document their data sources, data collection methods, and any preprocessing steps taken.
2. During Analysis:
    - In this stage, researchers should focus on conducting their analysis in a reproducible manner.
    - It is important to use scripts or code to perform data analysis, as this allows for transparency and easy replication of results.
3. After Analysis: Finalizing Results and Sharing:
    - Once the analysis is complete, researchers should finalize their results and make them accessible to others.
    - This includes sharing the data, code, program versions, parameters, and important intermediate results.

By following this three-step framework, researchers can enhance the reproducibility of their research and make it easier for others to validate and build upon their findings.

Three-step framework and Check-list guide for Reproducible Research

Step 1: Before data analysis

- ☐ Are raw data safely stored in multiple locations using multiple media?
- ☐ Are final data stored in a portable, non-proprietary format?
- ☐ Are final data formatted appropriately for analysis?
- ☐ Are data paired with adequate metadata?

Step 2: During data analysis

- ☐ Is code clean, readable, and appropriately formatted?
- ☐ Is code thoroughly commented?
- ☐ Have data and code been reviewed by at least one collaborator or friend?
- ☐ Have all software versions and computing environments been documented?

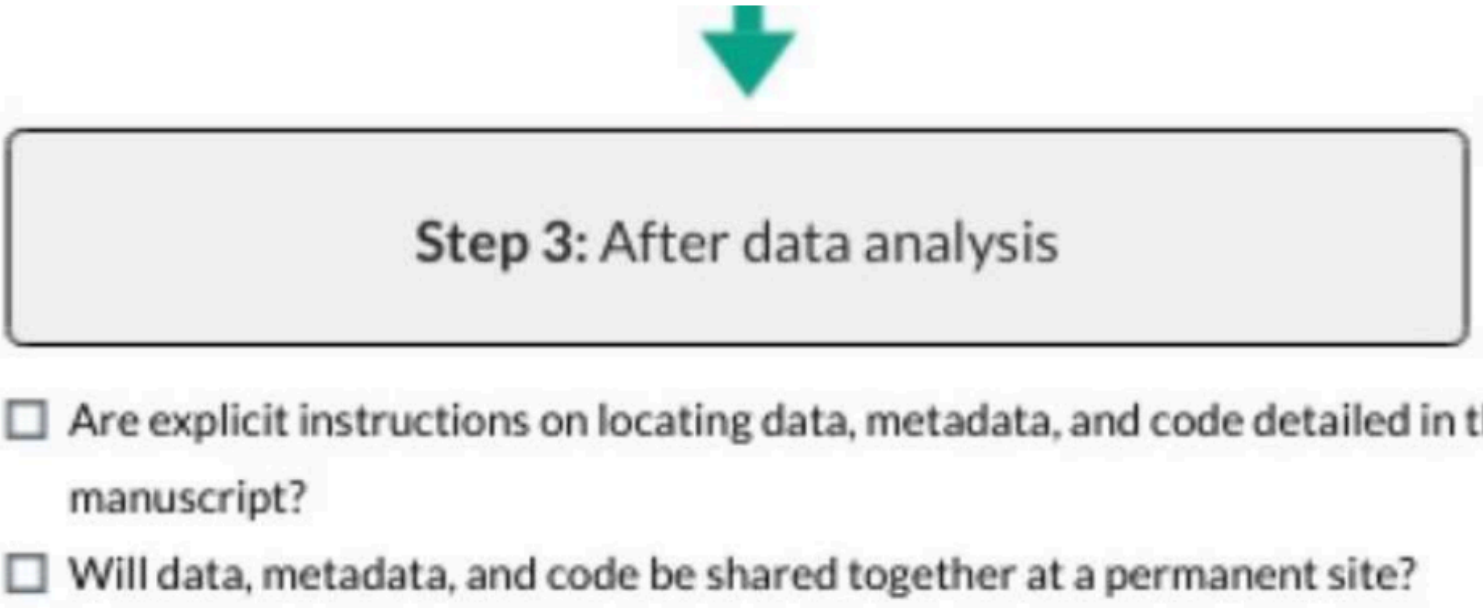


Fig. 1. A 10-point checklist to guide researchers toward greater reproducibility in their research. Researchers should give careful thought before, during, and after analysis to ensure reproducibility of their work.