CE5033 Statistical Methods and Data Mining

20240305 Exercise

1. Which of the following is a measure of variability?

   (A) Mode (B) Median (C) Mean (D) Standard Deviation

Reason: Standard deviation is a measure of variability or dispersion in a dataset, indicating how spread out the numbers are from the mean (average) of the dataset.

2. When you want to show the relationship between two continuous variables, what is the best type of figure to use?

   (A) Pie chart (B) Histogram (C) Scatter plot (D) Bar chart

Reason: A scatter plot is ideal for showing the relationship between two continuous variables, as it displays data points based on two variables, one on each axis, allowing for the observation of patterns, trends, or correlations.

3. What measure of central tendency is most affected by outliers?

   (A) Range (B) Mean (C) Mode (D) Median

Reason: The mean, being the average of a dataset, is most affected by outliers since it involves the sum of all values. A single extreme value can significantly skew the mean, unlike the median or mode, which are more robust to outliers.

4. Which concept is crucial for making informed decisions based on data?

   (A) Statistical thinking (B) Aesthetic design in visualization

   (C) Advanced calculus (D) Software engineering principles

Reason: Statistical thinking involves understanding and using statistical methods to analyze data, recognize patterns, and make informed decisions based on data analysis. It is crucial for making data-driven decisions.

5. Histograms differ from bar plots in that they are used for:

   (A) Categorical data (B) Continuous data

   (C) Data over time (D) Comparative data analysis

Reason: Histograms are used for continuous data to show the distribution of the data across different intervals or bins, allowing for the visualization of the shape, spread, and central tendency of the data.

6. What does a boxplot NOT typically show?

   (A) Median (B) Interquartile range (C) Mean (D) Outliers

Reason: A boxplot typically shows the median, interquartile range, and outliers but does not explicitly show the mean. The mean could be calculated separately and added to the boxplot if necessary.

7. If a company wants to visualize sales performance across different regions for each quarter, which figure would be least effective?

   (A) Line plot (B) Bar plot (C) Pie chart (D) Scatter plot

Reason: A pie chart would be the least effective for visualizing sales performance across different regions for each quarter as it is better suited for showing proportions of a whole at a single point in time rather than trends over time or comparisons across multiple categories.

8. What is the primary advantage of using a histogram over a bar plot for data visualization?

   (A) Histograms show distributions of continuous data variables

   (B) Histograms are better for categorical data

   (C) Histograms can only show proportions (D) Histograms are used for time series data

Reason: The primary advantage of using a histogram over a bar plot is that histograms are designed to show the distribution of continuous data variables, revealing patterns in the data such as normal distribution, skewness, or bimodality.

9. The process of selecting a subset of data from a larger dataset to analyze and draw conclusions is known as:

   (A) Bias (B) Sampling (C) Bootstrapping (D) Stratification

Reason: Sampling refers to the process of selecting a subset of data from a larger dataset to analyze and draw conclusions, which is fundamental in statistics for studying populations when it is impractical to examine the entire population.

10. What characteristic of data is best visualized using a box plot?

    (A) The mean value of the dataset (B) The relationship between two variables

    (C) The distribution and outliers in the dataset

    (D) The proportion of categories within the dataset

Reason: A box plot is specifically designed to visualize the distribution of data through its quartiles and also effectively highlights outliers, making it ideal for understanding the spread and identifying unusual observations in the dataset.

11. What is the primary benefit of using a pie chart in data visualization?

    (A) To show changes over time (B) To display the distribution of a continuous variable

    (C) To illustrate the proportion of categories within a whole

    (D) To compare the mean values of different groups

Reason: The primary benefit of using a pie chart is to display the proportions of categories

within a whole, making it easy to see at a glance how a total is divided among different categories.

12. What does the term “sampling bias” refer to?

(A) The preference for larger sample sizes in statistical studies

(B) The systematic error due to the unrepresentative sample selection

(C) The variability observed in sample statistics across different samples

(D) The tendency to include only positive outcomes in a sample

Reason: Sampling bias refers to the systematic error that occurs when the method of selecting a sample causes it to be unrepresentative of the population, potentially leading to biased outcomes in the analysis.

13. The bootstrap method is particularly useful for:

(A) Determining the sample size needed for an experiment

(B) Estimating the mean of a population without sampling

(C) Eliminating bias in sampled data

(D) Estimating the sampling distribution of a statistic or model parameters

Reason: The bootstrap method is particularly useful for estimating the sampling distribution of a statistic (like the mean or median) or model parameters by resampling with replacement from the original sample. This allows for the estimation of the variability or confidence intervals of the statistic without relying on normal distribution assumptions.

14. Which figure is preferred for visualizing the distribution of rental prices for apartments listed in a city?

(A) Bar plot (B) Line plot (C) Histogram (D) Pie chart

Reason: A histogram is preferred for visualizing the distribution of rental prices for apartments listed in a city because it can effectively show the distribution of continuous data, such as prices, across different intervals.

15. Which distribution is particularly useful for analyzing the rate of occurrence of rare events in a fixed space or time?

(A) Poisson (B) Normal (C) Binomial (D) Exponential

Reason: The Poisson distribution is particularly useful for analyzing the rate of occurrence of rare events in a fixed space or time, as it models the probability of a given number of events happening in a fixed interval.

16. Bootstrap method is for estimating:

(A) Bias (B) Mean (C) Sampling distribution (D) Population size

Reason: The bootstrap method is used for estimating the sampling distribution of a statistic,

allowing researchers to understand the variability or confidence intervals of the statistic based on the original sample.

17. Discrete data refers to values that are:

(A) Countable and take on specific values (B) Continuously varying across a range

(C) Measured in infinite decimal places (D) Not countable and cannot be measured

Reason: Discrete data refers to values that are countable and take on specific values. This contrasts with continuous data, which can take any value within a range and is measured rather than counted.

18. In the study of probability, what does a continuous random variable represent?

(A) A variable that can only take specific, separated values

(B) A variable that can take any value within a given range

(C) Only integer values (D) Outcomes of a Bernoulli process

19. What does the Central Limit Theorem describe?

(A) Data skewness (B) Sample mean distribution (C) Population size (D) Sample variance

Reason: The Central Limit Theorem describes how the distribution of the sample means will approximate a normal distribution as the sample size increases, regardless of the population's original distribution. This is fundamental for statistical inference, allowing for the use of normal probability calculations in many situations.

20. Poisson distribution models:

(A) Time between events (B) Success probability

(C) Number of events (D) Data spread

Reason: The Poisson distribution models the number of events occurring in a fixed interval of time or space, given a known constant mean rate. It is particularly useful for predicting the likelihood of a given number of occurrences within a specified interval, making it ideal for events that happen independently and at a constant average rate.