CE5033 Statistical Methods and Data Mining

20240312 Exercise

1. Why is sampling important in statistics?

   (A) It reduces research accuracy. (B) It increases study costs.

   (C) It allows for population inferences. (D) It applies only to large populations.

Reason: Sampling is crucial because it enables researchers to make informed conclusions about a population from a subset of its members, thereby making the research process both practical and efficient.

2. 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 occurs when the selection process results in a sample that is not representative of the population being studied, potentially leading to skewed results.

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

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

Reason: A pie chart is designed to show parts of a whole at a single point in time, making it ineffective for visualizing trends or changes over multiple time periods, such as quarterly sales performance across regions. It lacks the capacity to clearly convey the temporal progression that is essential for comparing performance over time, making it the least effective option among those listed for this specific visualization task.

4. Which of the following is not a requirement for applying the Central Limit Theorem (CLT)?

   (A) The sample observations must be independent.

   (B) The sample size should be sufficiently large.

   (C) The population from which the sample is drawn must be normally distributed.

   (D) Variances of the populations being sampled should be equal.

Reason: The CLT applies regardless of the population's distribution, provided the sample size is sufficiently large.

5. What does a boxplot NOT typically show?

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

Reason: A boxplot displays the median, quartiles, and outliers, but it does not typically show the mean.

6. What does the likelihood function measure?

(A) Direct p-value calculation. (B) Sample mean comparison.

(C) Probability of observing data. (D) Null hypothesis validity.

Reason: The likelihood function measures the probability of the observed data under various assumptions about the statistical model parameters, not the data's probability itself.

7. The Central Limit Theorem (CLT) facilitates statistical inference by:

(A) Allowing the use of small samples for any population.

(B) Guaranteeing that population distributions are normal.

(C) Enabling predictions about the distribution of sample means.

(D) Eliminating the variability in sample means.

Reason: The CLT allows for the approximation of the sampling distribution of the mean to be normal, regardless of the population distribution, facilitating the use of normal distribution in statistical inference.

8. Which statistical method estimates the distribution of a sample statistic through resampling?

(A) Central Limit Theorem (B) Standard deviation calculation

(C) Bootstrapping (D) Bias correction

Reason: Bootstrapping is a resampling technique used to estimate the distribution of a sample statistic by sampling with replacement from the data.

9. Which of the following is NOT a measure of central tendency?

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

Reason: The range is a measure of variability, not central tendency, which includes mean, median, and mode.

10. A histogram is most suitable for visualizing what type of data?

(A) Categorical data across different categories

(B) The relationship between two quantitative variables

(C) The distribution of a continuous data variable (D) Proportions of categories within a whole

Reason: Histograms are ideal for showing the distribution of continuous data, as they display the frequency of data points within specified ranges.

11. What does the standard error measure?

(A) Bias in the sample (B) Variability of a statistic (C) The mean difference (D) Sample size

Reason: The standard error measures the precision of a sample statistic (e.g., the mean) as an estimate of the population parameter, reflecting how much the statistic varies across different samples.

12. Interval estimation provides a range of values known as what?

(A) Point estimate (B) Variance estimate (C) Confidence interval (D) Probability distribution

Reason: Interval estimation results in a confidence interval, providing a range of values within which

the true population parameter is likely to lie.

13. What does interval estimation provide that point estimation does not?

(A) A single best estimate of a population parameter.

(B) A range of plausible values for a population parameter.

(C) A higher degree of bias in estimates. (D) Less accurate estimates of population parameters.

Reason: Unlike point estimation, which provides a single estimate, interval estimation offers a range, accounting for the uncertainty in the estimate.

14. What does the null hypothesis (H_0) usually state?

(A) No effect exists. (B) An effect exists.

(C) The sample mean is unique. (D) Population means differ.

Reason: The null hypothesis typically posits that there is no effect, difference, or relationship in the population, serving as a baseline for testing.

15. What does a p-value smaller than significance level (α) suggest?

(A) H_0 is true. (B) Insufficient evidence against H_0.

(C) Data align with H_1. (D) A larger sample is needed.

Reason: A p-value lower than the predetermined significance level suggests there is sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis.

16. How does inferential statistics differ from descriptive statistics?

(A) It only uses graphical methods to describe data.

(B) It makes predictions about a population based on sample data.

(C) It avoids using any form of data sampling. (D) It cannot be used for hypothesis testing.

Reason: Inferential statistics use sample data to make generalizations about a larger population, unlike descriptive statistics, which summarize the features of a dataset.

17. In hypothesis testing, a significant level of 0.05 means:

(A) 5% chance of Type II error (B) 5% of the data are outliers

(C) 5% chance the null hypothesis is true (D) 5% chance of Type I error

Reason: A significance level of 0.05 indicates a 5% risk of rejecting the null hypothesis when it is actually true.

18. A statistically significant result is indicated by a p-value that is:

(A) Higher than the significance level (α). (B) Greater than 0.5.

(C) Lower than or equal to the significance level (α). (D) Exactly equal to 1.

Reason: A p-value at or below the significance level suggests the observed data are unlikely under the null hypothesis, indicating statistical significance.

19. Estimation in statistical inference is used to:

(A) Infer population parameters from sample data. (B) Predict future sample data.

(C) Summarize data without generalization. (D) Apply only to normally distributed populations.

Reason: Estimation allows statisticians to use sample data to make inferences about unknown population parameters.

20. A Type I error occurs when:

(A) The null hypothesis is correctly rejected. (B) The null hypothesis is rejected when it is true.

(C) The alternative hypothesis is accepted when false. (D) A sample is biased.

Reason: A Type I error represents a false positive, where the null hypothesis is incorrectly rejected.