CE5033 Statistical Methods and Data Mining

20240319 Exercise

1. Which of the following is true regarding the Central Limit Theorem (CLT)?

   (A) The CLT requires the population from which samples are drawn to be normally distributed.

   (B) The CLT allows us to use the normal distribution to approximate the sampling distribution of the mean for large sample sizes.

   (C) The CLT states that the sample means will equal the population mean for large samples.

   (D) The CLT applies only to sample sizes smaller than 30.

Reason: The CLT states that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the population's distribution.

2. In hypothesis testing, what is the purpose of setting a significance level (α)?

   (A) To determine the sample size needed for the study.

   (B) To calculate the standard error of the mean.

   (C) To ensure the sample accurately represents the population.

   (D) To control the probability of making a Type I error.

Reason: Setting a significance level (α) defines the threshold for the probability of rejecting the null hypothesis when it is actually true (Type I error).

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

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

4. A researcher conducts a hypothesis test and calculates a p-value of 0.03. If the significance level is set at 0.05, what should the researcher conclude?

   (A) Fail to reject the null hypothesis, as the p-value is greater than 0.05.

   (B) Reject the null hypothesis, as the p-value is less than 0.05.

   (C) Increase the sample size and conduct the test again.

   (D) There is a 3% chance the null hypothesis is true.

Reason: A p-value lower than the significance level (α) indicates that the observed data are unlikely under the null hypothesis, leading to its rejection.

5. How does increasing the sample size affect the standard error of the mean in the context of CLT?

   (A) It increases proportionally with the sample size.

   (B) It remains unchanged regardless of the sample size.

   (C) It decreases as the sample size increases.

   (D) It initially decreases but then increases with larger sample sizes.

Reason: In the context of the Central Limit Theorem (CLT), the standard error (SE) of the mean is defined as the standard deviation of the population (σ) divided by the square root of the sample size (n), i.e., SE = $\sigma/\sqrt{n}$. As the sample size increases, the denominator of this fraction becomes larger, resulting in a smaller standard error. This decrease in the standard error indicates that the sample mean becomes a more precise estimator of the population mean as the sample size grows.

6. Effect size is a measure of:

(A) Significance level (B) Data variability (C) Difference magnitude (D) Type I error probability

Reason: Effect size measures the magnitude of the difference between groups or the strength of a relationship, providing insight beyond statistical significance.

7. What is the null hypothesis a statement of?

(A) No effect or difference (B) Significant effect (C) Expected outcome (D) Hypothesis to prove

Reason: The null hypothesis typically posits that there is no effect, difference, or relationship between groups or variables being studied.

8. 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 variability or dispersion of a sample statistic (e.g., sample mean) from the true population parameter.

9. A one-sample t-test compares:

(A) Single group mean to known mean (B) Means of two independent groups

(C) Means of two related groups (D) Variance within a single group

Reason: A one-sample t-test compares the mean of a single sample group to a known or hypothesized population mean.

10. A Type I error occurs when:

(A) A test fails to detect a true effect. (B) The null hypothesis is incorrectly accepted.

(C) The null hypothesis is falsely rejected. (D) The p-value exceeds the significance level.

Reason: A Type I error occurs when the null hypothesis is incorrectly rejected despite being true.

11. Which of the following best describes a Type II error in hypothesis testing?

(A) Rejecting the null hypothesis when it is true.

(B) Failing to reject the null hypothesis when it is false.

(C) Accepting the alternative hypothesis when it is false.

(D) Conducting a test without predefining a significance level.

Reason: A Type II error occurs when the test fails to reject the null hypothesis even though the alternative hypothesis is true.

12. The sampling distribution describes:

(A) A single sample's variability (B) Population variability

(C) The variability of a statistic across samples (D) The likelihood of sampling bias

Reason: The sampling distribution describes how a sample statistic, like the mean, varies from sample to sample.

13. Bootstrapping is used to:

(A) Replace traditional sampling methods (B) Estimate the distribution of a statistic

(C) Eliminate sampling bias (D) Reduce the need for a sample

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

14. A two-tailed test is appropriate when:

(A) The direction of the effect is not specified (B) Only positive effects are considered

(C) The sample size is below 30 (D) Variance is known

Reason: A two-tailed test is used when researchers are interested in determining whether there is a difference in either direction, without specifying the direction of the effect.

15. Which statistical test would you use to compare the means of two related samples?

(A) Independent samples t-test (B) Paired samples t-test

(C) One-sample t-test (D) Chi-square test

Reason: The paired samples t-test is used to compare the means of two related samples, such as measurements taken from the same group at two different times.

16. What is the primary purpose of hypothesis testing in statistics?

(A) To describe the characteristics of the sample

(B) To prove that the null hypothesis is true

(C) To make inferences about population parameters based on sample statistics

(D) To estimate the exact value of population parameters

Reason: Hypothesis testing is used to determine whether there is enough evidence in a sample to infer that a certain condition holds for the entire population.

17. The p-value is:

(A) The probability that the null hypothesis is true.

(B) The probability of observing the test results, or more extreme, under the null hypothesis.

(C) A measure of the effect size.

(D) Determined after deciding whether to reject the null hypothesis.

Reason: The p-value quantifies the evidence against the null hypothesis, indicating how likely the observed data (or more extreme) would be if the null hypothesis were true.

18. A 95% CI means:

(A) 95% chance true parameter is within (B) 95% sample data fall within

(C) 95% CIs will contain true parameter (D) Parameter has 5% chance not in interval

Reason: A 95% confidence interval means that if we were to take many samples and construct a CI from each, we'd expect 95% of those intervals to contain the true population parameter.

19. Statistical power is the probability the test correctly:

(A) Accepts H_0 when true (B) Rejects H_0 when H_A is true

(C) Identifies no effect when none (D) Making a Type I error

Reason: Statistical power is the probability that the test correctly rejects the null hypothesis when the alternative hypothesis is true, minimizing Type II errors.

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

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

Reason: The mean is sensitive to outliers as it takes into account the value of every data point, so extreme values can significantly affect it.