CE5033 Statistical Methods and Data Mining

20240402 Exercise

1. Inferential statistics is distinct from descriptive statistics due to its ability to:

   (A) Use only graphical data descriptions.

   (B) Estimate population parameters based on sample data.

   (C) Avoid all forms of sampling.

   (D) Avoid conducting hypothesis testing.

Reason: Inferential statistics use data from a sample to make inferences or predictions about a population, unlike descriptive statistics, which only describe the data at hand.

2. Which statement correctly describes the Central Limit Theorem (CLT)?

   (A) If the sample size is large enough, the CLT states that the sampling distribution of the sample mean will be uniformly distributed.

   (B) The CLT states that for any given sample size, the mean of the sample will always be equal to the mean of the population.

   (C) The CLT states that as the sample size increases, the variance of the sampling distribution of the sample mean decreases.

   (D) The CLT states that for sufficiently large sample size, the sampling distribution of the sample means will approximate a normal distribution regardless of the distribution of the population.

Reason: The CLT allows for the approximation of the sampling distribution of the sample mean to a normal distribution as the sample size increases, irrespective of the population's original distribution.

3. Which scenario correctly describes a Type I error?

   (A) Correct rejection of a false null hypothesis.

   (B) Presence of bias in the sample selection.

   (C) Acceptance of a false alternative hypothesis.

   (D) Rejection of a true null hypothesis.

Reason: A Type I error occurs when the null hypothesis is true, but the decision is incorrectly made to reject it, usually based on the p-value being smaller than the chosen significance level.

4. In point estimation, the standard error of the mean primarily measures:

   (A) The degree to which the point estimate is larger than the population mean.

   (B) The bias of the point estimate in estimating the population mean.

   (C) The precision of the point estimate in estimating the population mean.

   (D) The probability that the point estimate is correct.

Reason: The standard error of the mean quantifies how far the sample mean is likely to be from the population mean, thereby indicating the precision of the point estimate.

5. Effect size in statistical analysis is important because it:

   (A) Provides a measure of the practical significance of the study findings.

(B) Determines the statistical significance of the results.

(C) Reduces the probability of committing Type I and Type II errors.

(D) Indicates the adequacy of the sample size for the analysis.

Reason: While statistical significance (e.g., p-value) indicates whether an effect exists, the effect size quantifies the magnitude of the difference or relationship, providing insight into its practical importance.

6. Which of the following statements is true regarding the significance level (α) in hypothesis testing?

(A) It is the probability of making a Type II error.

(B) It determines the threshold for the power of the test.

(C) It is the probability that the sample data occurred by random chance.

(D) It is the predetermined threshold at which the null hypothesis is rejected.

Reason: The significance level (α) is set before analyzing data as a threshold for determining whether the observed data are unlikely enough under the null hypothesis to warrant its rejection.

7. In hypothesis tests, the significance level (α) is typically set at 0.05. This means:

(A) An effect size of 5% in the population parameters is guaranteed.

(B) The null hypothesis is true with a probability of 95%.

(C) There is a 5% chance of type I error due to rejection of a true null hypothesis.

(D) Only when the p-value is greater than 0.05 can the null hypothesis be rejected.

Reason: Setting α at 0.05 means the researcher is willing to accept a 5% risk of rejecting the null hypothesis when it is actually true (Type I error).

8. In evaluating the effectiveness of a new drug, what type of error is made if researchers conclude that the drug is effective when it is not?

(A) Sampling error (B) Type I error (C) Measurement error (D) Type II error

Reason: Concluding that a treatment has an effect when it does not is a Type I error, where a true null hypothesis is incorrectly rejected.

9. In hypothesis testing, what does a two-tailed test tell us about the research hypothesis?

(A) The statement specifies a direction, either greater than or less than.

(B) The test does not specify a direction but rather detects any differences.

(C) Only negative differences are evaluated in the test.

(D) Only positive differences are evaluated in the test.

Reason: A two-tailed test is used when the researchers are interested in determining if there is a difference, regardless of direction, between two groups or treatments.

10. In a customer feedback form, respondents are asked to check one of the boxes marked as 'poor,' 'fair,' 'good,' 'very good,' and 'excellent' to rate service quality. This type of data collection is best described as collecting:

(A) Ordinal data (B) Nominal data (C) Interval data (D) Ratio data

Reason: These categories represent ordered rankings without implying a specific numerical distance

between each rating, characteristic of ordinal data.

11. What is the relationship between the confidence level and the width of a confidence interval?
    (A) The confidence level has no impact on the width of the confidence interval.
    (B) Higher confidence levels result in narrower confidence intervals.
    (C) Higher confidence levels result in wider confidence intervals.
    (D) Only the sample size affects the width of the confidence interval, not the confidence level.

Reason: A higher confidence level means that we require a wider interval to be more certain that it contains the population parameter, reflecting increased uncertainty.

12. The power of a statistical test is defined as:
    (A) The probability that the null hypothesis is correctly accepted.
    (B) The probability of detecting a significant result when there is no actual effect.
    (C) The probability of committing a type I error.
    (D) The probability that the null hypothesis is correctly rejected when it is false.

Reason: Power is the likelihood of detecting an effect or difference when one truly exists, which means correctly rejecting a false null hypothesis.

13. Why would a researcher use bootstrap method instead of traditional parametric inference methods?
    (A) When the researcher is unsure whether data meet parametric assumptions.
    (B) Only when the sample size is too large for traditional methods to handle.
    (C) Bootstrap methods are only used when the data are normally distributed.
    (D) It is only applicable for estimating the mean of a sample.

Reason: Bootstrap methods do not rely on the assumption of normality or other specific distributional forms, making them useful when data may not meet the assumptions required for parametric tests.

14. What does "resampling with replacement" mean in bootstrapping?
    (A) Each selected unit is returned to the record before the next draw, meaning it could potentially be selected again.
    (B) Each sample drawn is kept separate from the dataset to preserve independence.
    (C) Samples are drawn from the dataset without any chance of selecting the same unit more than once.
    (D) The dataset is partitioned into several non-overlapping samples.

Reason: In bootstrapping, "with replacement" allows the same observation to be included multiple times in the same resample, mirroring the process of sampling from an infinite population.

15. Which null and alternative hypotheses correctly represent a scenario where a researcher is testing whether a new drug (μ_{new}) is more effective than the existing standard ($\mu_{standard}$) in a one-tailed hypothesis test?
    (A) H_0: $\mu_{new} = \mu_{standard}$ versus H_1: $\mu_{new} \neq \mu_{standard}$
    (B) H_0: $\mu_{new} \neq \mu_{standard}$ versus H_1: $\mu_{new} = \mu_{standard}$
    (C) H_0: $\mu_{new} = \mu_{standard}$ versus H_1: $\mu_{new} > \mu_{standard}$

(D) H_0: $\mu_{new} > \mu_{standard}$ versus H_1: $\mu_{new} = \mu_{standard}$

Reason: This set of hypotheses is appropriate for a one-tailed test where the interest is specifically in whether the new drug is more effective, not just different.

16. When using a hypothesis test, why would a researcher choose to use a two-tailed test rather than a one-tailed test?

(A) To improve the chances of finding a result that is statistically significant or not.

(B) When the direction of the effect is not known or when it's important to detect effects in either direction.

(C) Only when the population standard deviation is unknown.

(D) Two-sided tests are only chosen when dealing with nominal data.

Reason: A two-tailed test is used when researchers are interested in differences regardless of direction, allowing for the detection of an effect whether it's positive or negative.

17. What is the most appropriate scenario for conducting a one-tailed hypothesis test?

(A) Testing whether a new drug is different in its effectiveness from an existing drug.

(B) Determining whether a new teaching method results in higher or lower test scores compared to the traditional method.

(C) Assessing whether a dietary supplement decreases blood cholesterol levels.

(D) Comparing the mean annual income of two professions to determine if there is a difference.

Reason: This scenario implies a specific direction of interest (decrease), making it suitable for a one-tailed test where the alternative hypothesis is directional.

18. When conducting a hypothesis test regarding a population mean, if the test statistic falls outside the rejection region, you should:

(A) Reject the null hypothesis.

(B) Fail to reject the null hypothesis.

(C) Increase the sample size and recalculate the test statistic.

(D) Convert the test into a one-sided test for a stronger conclusion.

Reason: If the test statistic does not fall into the critical region (the region beyond the critical values for rejecting the null hypothesis), there is insufficient evidence to reject the null hypothesis.

19. In hypothesis testing, what is the consequence of setting a very low significance level (e.g., $\alpha = 0.01$) compared to a more conventional level (e.g., $\alpha = 0.05$)?

(A) It increases the chance of making a Type I error.

(B) It decreases the chance of making a Type II error.

(C) It makes it more difficult to reject the null hypothesis.

(D) It makes the confidence interval narrower.

Reason: A lower significance level means a stricter criterion for rejecting the null hypothesis, requiring stronger evidence (a smaller p-value) to do so.

20. Which statement accurately reflects the relationship between hypothesis tests and confidence intervals?

(A) A hypothesis test is used to determine if a confidence interval contains the true population parameter.

(B) If a confidence interval for a mean difference includes zero, the null hypothesis of no difference cannot be rejected at the corresponding significance level.

(C) The purpose of confidence intervals is solely for the determination of the sample size necessary for hypothesis testing.

(D) Hypothesis tests provide a range of values that may include the population parameter, whereas confidence intervals do not.

Reason: The presence of the null value within the confidence interval suggests that at the given confidence level, there is insufficient evidence to reject the null hypothesis, highlighting the intrinsic link between confidence intervals and hypothesis testing outcomes.