Computing For Medicine
CSE585/BIO546
End Semester Examination

Instructions

1. Exam Duration: 1.5 hour
2. Max Marks: 50
3. Only one option in MCQs is correct. MCQs have to be submitted on the Answer Sheet, responses marked on the Question Paper will NOT be considered.

Name:

Roll No:

Ques 1. A group of researchers is investigating the link between smoking and lung cancer. They follow two groups of participants for 10 years:

- Group 1: Participants who were smokers at the start of the study.
- Group 2: Participants who were non-smokers at the start of the study.

Over the 10 years, the researchers record which participants develop lung cancer. They then compare the incidence of lung cancer between the two groups to assess the risk associated with smoking. What type of study design does this represent? (2)

A. Retrospective Cohort Study
B. Observational Cross-Sectional Study
C. Prospective Cohort Study
D. Randomized Controlled Trial

Ques 2. Which of the following best explains why BERT uses bidirectional context in its model architecture? (2)

A. To speed up training through parallel processing.
B. To encode both past and future information in word representations.
C. To improve memory efficiency in large-scale language tasks.
D. To simplify model interpretability.

Ques 3. In the context of hypothesis testing, if the p-value of a test is 0.03 at a significance level of 0.05, which of the following conclusions is CORRECT? (2)

A. The null hypothesis is rejected with 95% confidence.

B. The alternative hypothesis is accepted with 97% confidence.
C. The null hypothesis is rejected with 97% confidence.
D. The test does not provide sufficient evidence to reject the null hypothesis.

Ques 4. A healthcare organization is designing a meta-model to integrate various processes. The meta-model includes:

- Chronological workflows for patient care.
- Data models for health records.
- Guidelines for interoperability between systems.

What is the key advantage of using a meta-model in this scenario?

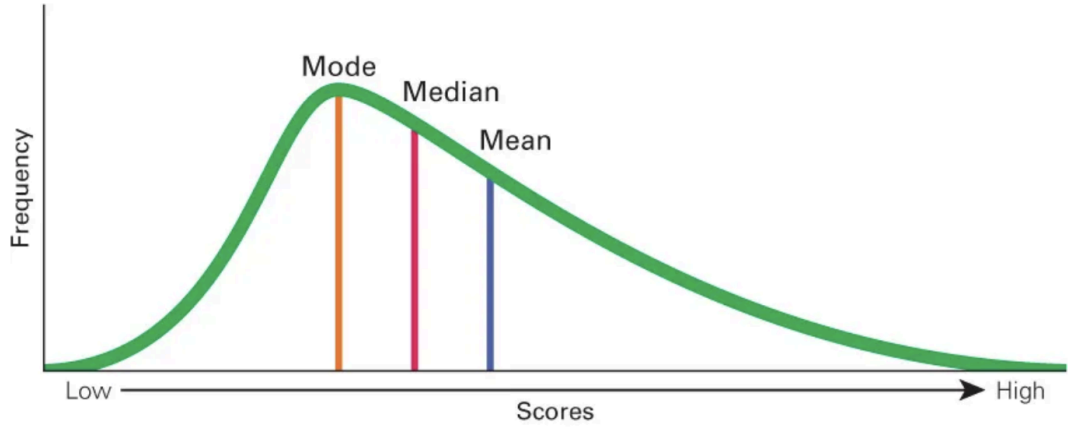
A. It ensures the health system uses only one type of database.
B. It standardizes the design and interaction of different models for seamless integration.
C. It focuses solely on technical aspects, ignoring workflows and processes.
D. It replaces all existing models with a single comprehensive model.

Ques 5. Which scenario best illustrates a Type I error in a healthcare context? (2)

A. Accepting the null hypothesis when it is true
B. Failing to diagnose a diseased patient
C. Incorrectly concluding a treatment has no effect when it does
D. Diagnosing a healthy patient as diseased

Ques 6. Briefly answer the following (in 2-3 sentences): (2*5 = 10)

A. Explain the purpose of residual connections in transformer models.
B. The graph below shows the distribution of a dataset. What type of skewness does this distribution represent? Explain how the positions of the mean, median, and mode indicate the skewness.



C. What is the significance of feed-forward layers in the transformer architecture
D. How can an incorrect α level lead to negative consequences in healthcare decisions. Explain with an example.
E. What is the purpose of applying dropout in transformer models

Ans.

A. **Residual connections in transformers**:
Residual connections help prevent gradient vanishing issues and allow deeper networks to train efficiently by enabling gradient flow directly through layers.
B. **Skewness in the graph**: Positively skewed/ right-skewed
If the mean > median > mode, the distribution is **right-skewed** (positively skewed). If the mean < median < mode, it is **left-skewed** (negatively skewed). Skewness occurs due to outliers pulling the mean.
C. **Significance of feed-forward layers in transformers**:
Feed-forward layers perform non-linear transformations on attention outputs, enhancing the model's ability to learn complex representations.
D. **Impact of incorrect α\alphaα**:
Setting α\alphaα too low may fail to detect meaningful effects (Type II error). For instance, in drug trials, an overly stringent α\alphaα could reject a potentially effective treatment.
E. **Purpose of dropout in transformers**:
Dropout reduces overfitting by randomly deactivating neurons during training, forcing the network to generalize better.

Ques 7. A hospital is conducting several studies to evaluate the effects of different treatments and conditions on patient outcomes. Based on the scenarios below, identify which statistical test should be used and briefly justify your choice: (2*5 = 10)

1. **Scenario 1:** A dietician measures the mean cholesterol level of a group of 50 patients and wants to test if it differs significantly from the national average of 200 mg/dL.
2. **Scenario 2:** Two groups of patients (30 each) are tested for blood pressure reduction after being given two different drugs. The goal is to compare the mean blood pressure reductions between the two groups.
3. **Scenario 3:** A researcher measures the blood glucose levels of 20 diabetic patients before and after a new insulin therapy to determine if there is a significant change.
4. **Scenario 4:** A study compares the pain relief scores from three different medications in 60 patients (20 patients per medication).

5. **Scenario 5:** A small clinic collects satisfaction ratings (on a 5-point scale) for two different therapy sessions. The data does not meet normality assumptions, and the clinic wants to test if the ratings differ significantly between the two therapies.

Ans.

Scenario 1: One-sample t-test

- Testing if the mean cholesterol level differs from a known value.

Scenario 2: Independent two-sample t-test

- Comparing means between two independent groups (different drugs).

Scenario 3: Paired t-test

- Comparing pre-treatment and post-treatment glucose levels in the same group.

Scenario 4: ANOVA

- Comparing means across more than two groups (three medications).

Scenario 5: Wilcoxon Rank-Sum Test

- Non-parametric test for two independent groups with non-normal data.

Ques 8. A national health program is developing an integrated platform to connect hospitals, clinics, and labs for seamless data exchange and analysis. The system needs to:

- Record diagnoses in a standardized format to ensure consistent interpretation across healthcare facilities.
- Enable real-time exchange of patient observations, such as lab test results, with minimal technical overhead.

Answer the following:(5)

1. Why is it important to use a standardized vocabulary for recording diagnoses in such a system? (2)
2. Describe how a structured data exchange format can improve communication between hospitals and labs compared to traditional file-sharing methods. (2)
3. What are the potential challenges in implementing such a system, and how could they be addressed? (1)

Ans.

1. **Importance of standardized vocabulary (2 marks):**
   A standardized vocabulary ensures consistent interpretation and reduces ambiguity in diagnoses across different healthcare facilities.
2. **Benefits of structured data exchange format (2 marks):**
   Structured formats enable automated parsing, reduce errors, and facilitate real-time data sharing compared to unstructured file sharing.
3. **Challenges and solutions (1 mark):**
   Challenges include system interoperability and high implementation costs. Solutions involve adopting global standards and phased integration.

Ques 9. Explain the concept of confounding in statistical analysis. How can confounding factors influence the results of correlation analysis in healthcare data? Explain with an example. (5)

Ans. Confounding occurs when an external variable (a **confounder**) influences both the independent variable and the dependent variable, potentially distorting the perceived relationship between them and creating a false or biased association. This can lead to incorrect conclusions about causality or the strength of a relationship.

Example:
In studying smoking and lung cancer, age might act as a confounder, as older individuals are both more likely to smoke and more prone to cancer. Proper statistical controls, like multivariable regression, are needed to mitigate confounding effects.

Ques 10. A pharmaceutical company has developed a rapid diagnostic test for a rare but serious infectious disease. The disease affects 1 in 10,000 individuals in the general population. The test has a sensitivity of 99% and a specificity of 95%. During a community health drive, 10,000 individuals are tested, and your test result comes back positive. What is the probability that you actually have the disease? How many false positives would you expect among the 10,000 individuals tested? (10)

- Sensitivity ($P(T^+|D)$) = 0.99
- Specificity ($P(T^-|D')$) = 0.95
- Prevalence ($P(D)$) = 0.0001
- Probability of no disease ($P(D')$) = 1 - $P(D)$ = 0.9999

Calculate $P(T^+)$:

$$P(T^+) = P(T^+|D)P(D) + P(T^+|D')P(D')$$

$$P(T^+) = (0.99)(0.0001) + (1 - 0.95)(0.9999) = 0.000099 + 0.049995 = 0.050094$$

Posterior probability:

$$P(D|T^+) = \frac{P(T^+|D)P(D)}{P(T^+)} = \frac{(0.99)(0.0001)}{0.050094} \approx 0.001977\,(0.2\%)$$

2. **False positives (5 marks):**

Number of false positives = $(1 - \text{Specificity}) \times \text{Non-diseased individuals}$

$$\text{False positives} = (1 - 0.95) \times 9999 = 0.05 \times 9999 \approx 500$$