

COMPUTING FOR MEDICINE

QUIZ 3

Name and Roll No:

Time: 45 minutes

Instructions: Some questions may have more than one correct answer. You need to select *all* the correct options to receive full credit.

Ques 1. What is/are the primary assumption(s) of a one-sample t-test? (5 marks)

- a) The sample data are independent.**
- b) The data come from a normally distributed population.**
- c) The population standard deviation is unknown.**
- d) The sample size must exceed 30.

Ques 2. Which of the following is/are not captured by correlation(s)? (5 marks)

- a) Hidden confounding factors.**
- b) Nonlinear relationships.**
- c) Linear trends.
- d) None of the above.

Ques 3. Which of the following concept(s) apply to the training and functionality of PEGASUS and BioBERT? (5 marks)

- a) PEGASUS uses a Gap Sentence Generation (GSG) strategy for summarization tasks.**
- b) BioBERT enhances general transformer models with domain-specific biomedical corpora.**
- c) Both PEGASUS and BioBERT rely on autoregressive training objectives.
- d) PEGASUS excels in abstractive summarization for general and domain-specific tasks.**

Ques 4. Which statement(s) correctly describe the Central Limit Theorem (CLT)?

- a) The distribution of sample means approaches normality as the sample size increases, regardless of the population distribution.**
- b) The Central Limit Theorem applies only to normally distributed populations.
- c) Larger sample sizes improve the approximation of the sample mean to a normal distribution.**
- d) CLT ensures that individual data points in the population will follow a normal distribution.

Ques 5. What is linear regression? [4] Under what circumstances would you choose to use simple linear regression, multiple linear regression, or linear regression with interaction effects? [4] Provide examples for each case in the context of healthcare data analysis [2]. (10 marks)

Ans. Simple linear regression is a statistical method used to model and analyze the relationship between two variables: one independent variable (X) and one dependent variable (Y). It assumes a linear relationship, where changes in the independent variable correspond to proportional changes in the dependent variable.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Where:

- Y : Dependent variable (outcome).
- X : Independent variable (predictor).
- β_0 : Intercept, the value of Y when $X = 0$.
- β_1 : Slope, the change in Y for a one-unit change in X .
- ϵ : Error term, accounting for variability not explained by X .

Assumptions

- Linearity: The relationship between X and Y is linear.
- Independence: Observations are independent of each other.

- Homoscedasticity: The variance of residuals (errors) is constant across all values of X.
- Normality of Errors: The residuals (errors) follow a normal distribution.
- No Multicollinearity: Since there's only one independent variable, multicollinearity isn't relevant in simple linear regression.

Circumstances for Using Different Types of Linear Regression (4 Marks)

a) Simple Linear Regression

Used when there is one independent variable.

Example in Healthcare: Predicting a patient's blood pressure (Y) based on age (X).

b) Multiple Linear Regression

Used when there are multiple independent variables.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

Example in Healthcare: Analyzing the impact of age, BMI, and cholesterol levels (X_1, X_2, X_3) on blood pressure (Y).

c) Linear Regression with Interaction Effects

When multiple independent variables are correlated (used to explore whether the effect of one variable depends on another). Interaction terms ($X_1 \times X_2$) are included in the model.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + \epsilon$$

Example:

Predicting blood pressure based on age (X_1) and whether a patient is on medication (X_2):

$$BP = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Medication}) + \beta_3(\text{Age} \cdot \text{Medication}) + \epsilon$$

Ques 6. Discuss the role of hypothesis testing in medical data science [3]. Explain the different types of errors & their implications [4]. Illustrate with examples how visualization tools (like box plots/histograms) can complement hypothesis testing in healthcare research [3] (10 marks)

Ans. Hypothesis testing is a method to make inferences about populations using sample data. It involves defining:

1. Null Hypothesis (H_0): Represents the default assumption (e.g., no effect or no difference).
2. Alternative Hypothesis (H_a): Represents the opposing claim (e.g., a significant effect or difference exists).

Through statistical tests, hypothesis testing evaluates whether the observed results in a sample are due to random chance or reflect actual effects in the population.

Different types of error are Type I Error and Type II Error

Type I error, also known as a **false positive**. It occurs when the null hypothesis is rejected based on the sample data, even though it is actually true in the population.

Example in Healthcare:

Suppose a clinical trial tests whether a new drug (Drug A) is more effective than a placebo.

- H_0 : Drug A has no effect on the disease compared to the placebo.
- H_a : Drug A is effective.

If the trial mistakenly rejects H_0 , the researchers might conclude that Drug A works when it does not. This could lead to:

Implications:

- Patients being prescribed an ineffective treatment.
- Increased healthcare costs due to unnecessary production and use of the drug.
- Ethical concerns, as patients might experience side effects without any benefit.

Type II error, also known as a **false negative**, occurs when the null hypothesis is not rejected based on the sample data, even though it is actually false in the population. In other words, it's a failure to recognize a real effect or difference.

Example in Healthcare:

In a study evaluating the effectiveness of a new cancer treatment:

- H_0 : The new treatment is no better than the standard treatment.
- H_a : The new treatment is better.

If the study fails to reject H_0 when H_a is true, the new treatment's benefits go unnoticed.

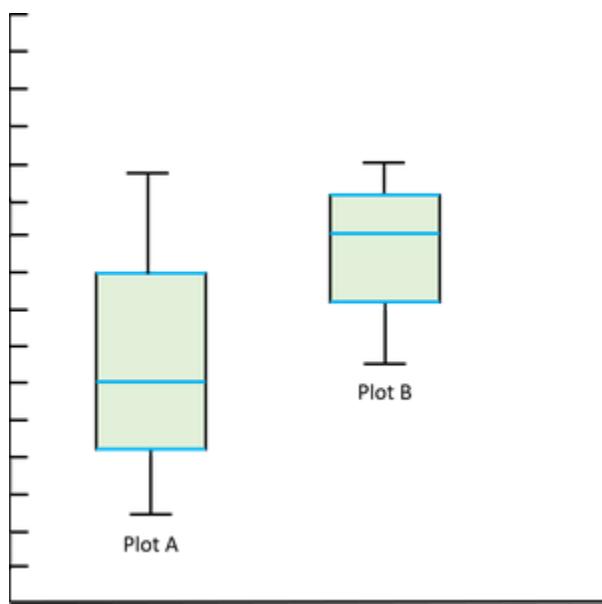
Implications:

- Delayed adoption of a potentially life-saving treatment.
- Missed opportunities for improving patient outcomes.
- Continued reliance on less effective treatments.

Visualization Tools Complementing Hypothesis Testing (3 Marks)

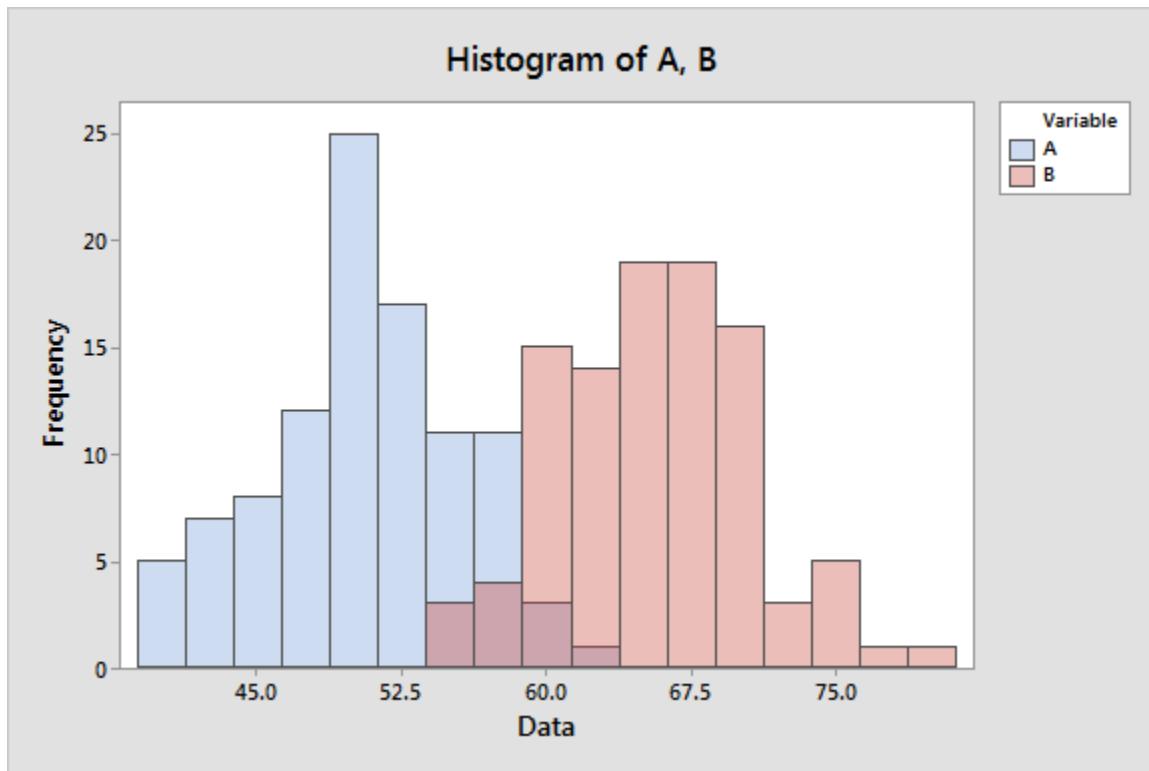
Box Plots: Useful for visualizing the spread and central tendency of data, making it easier to compare groups before performing a test.

Example: Comparing blood pressure distributions between treatment and control groups.



- Compare the Medians- Here the Median line of the plot B lies outside the box of Plot A.
- Compare the Dispersion or Spread of data- Here Plot A has a longer length than Plot B which means that the dispersion of data is more in plot A as compared to plot B. The length of whiskers also gives an idea of the overall spread of data. The extreme values (minimum &maximum) give the range of data distribution. Larger the range more scattered the data. Here Plot A has a larger range than Plot B.

- Comparing Outliers- The outliers give the idea of unusual data values which are distant from the rest of the data. More number of Outliers means the prediction will be more uncertain.
- Compare Skewness- Here Plot A is Positive or Right Skewed and Plot B is Negative or Left Skewed.



Example Application in Healthcare

Suppose the histogram represents cholesterol levels (variables A and B) in two patient groups:

- Group A: Patients following a diet plan.
- Group B: Patients not following a diet plan.

Hypothesis Testing Goal:

To test whether the diet plan significantly reduces cholesterol levels.

- A histogram like this provides an initial visual understanding of differences before applying statistical tests.

Using the Histogram in Hypothesis Testing

- Checking for Normality

- Identifying Variability
- Testing for Differences