



COMSATS UNIVERSITY ISLAMABAD (CUI)

DEPARTMENT OF COMPUTER SCIENCE
 SPRING 2025
 TERMINAL EXAM

Course: DSC300-Statistical Methods in Data Science
 Maximum Marks: 50
 Class: BSDS-4

Instructor: Basharat Mahmood
 Dated: July 03, 2025.

Table 1 shows a partial dataset that contains health and lifestyle information, including age, BMI, daily sunlight exposure, and vitamin D levels (scaled), with an indicator showing whether their vitamin D level is within the normal range.

Table 1: Sample data for Vitamin D deficiency prediction

Person	Age (yrs)	BMI ($\times 10^{-1}$)	Sunlight (hrs/day)	Vitamin D Level (ng/mL $\times 10^{-1}$)	Normal Vitamin D (0/1)
A	25	2.2	3.0	3.5	1
B	45	2.85	1.0	1.8	0
C	30	2.6	2.0	2.5	0
D	35	2.45	2.5	2.8	0
E	50	3	0.5	1.5	0
F	28	2.3	3.5	3.8	1
G	40	2.7	1.5	2.2	0
H	32	2.5	2.8	3	1

CLO2: Apply appropriate statistical methods to make inference about population parameters using sample statistics.

Question No. 1

15 Marks

- It is claimed that individuals with **more than 2.5 hours/day of sunlight exposure** (i.e., Sunlight > 0.25 in scaled data) have **higher average vitamin D levels** than those with **2.5 hours or less** (Sunlight ≤ 0.25). At the **0.05 level of significance**, test this claim using an appropriate **one-tailed two-sample test**. [5 Marks]
- Various studies claim that at least 75% of individuals who are **older than 30 years** and have **less than 2.5 hours of sunlight exposure per day** suffer from vitamin D deficiency (encoded as 0). Using the given dataset and a significance level of 0.05, test this assertion by conducting a one-sample z-test for proportion, provided the sample size is large enough for the normal approximation to be valid. [5 Marks]
- In part 2, you tested the claim that at least 75% of individuals over 30 years old with less than 2.5 hours/day of sunlight exposure are vitamin D deficient. State and explain what a Type I error and a Type II error would represent in this context. [3 Marks]
- Define the terms alpha (α) and beta (β) in the general context of hypothesis testing. [2 Marks]

CLO 3: Apply appropriate statistical models to analyse relationship among different variables.

Question No. 2

25 Marks

1. Using the provided dataset, examine whether there is a relationship between an individual's **Body Mass Index (BMI)** and their **Vitamin D status** (deficient = 0, normal = 1). Specifically:
 - a. Calculate the appropriate **correlation coefficient** between BMI and the binary variable indicating vitamin D deficiency. [4 Marks]
 - b. Based on the value of the correlation coefficient, interpret the **strength** and **direction** of the relationship. [2 Marks]
 - c. Calculate a **96% confidence interval** for the population correlation coefficient ρ . [4 Marks]
2. Using the provided dataset, construct a **multiple linear regression model** to predict an individual's **Vitamin D Level** based on the following two predictor variables: **BMI** and **Sunlight Exposure**
 - a. Fit a multiple linear regression model: $Y = a + b_1 \cdot \text{BMI} + b_2 \cdot \text{Sunlight}$ [7 Marks]
 - b. What does each coefficient tell you about the relationship between the predictor and vitamin D level, holding the other variable constant? [2 Marks]
 - c. Using your model, estimate the **Vitamin D Level** for a person with: BMI = 2.6 and Sunlight = 0.30 [1 Mark]
3. Identify purposes, interpretations, and key differences between correlation and regression analysis. [5 Marks]

CLO 4: Apply decomposition techniques to analyse time-series data.

Question No. 3

10 Marks

1. Identify the various components of variation in time series data. Additionally, describe the process of time series decomposition and how it helps analyze these components? [5 Marks]
2. Find the trend-values for the following data using the semi-average method. [5 Marks]

Quarter	2001	2002	2003	2004
Q1	500	550	530	570
Q2	480	510	490	520
Q3	600	650	620	660
Q4	700	740	710	750

The End