Time: 2 hr

Mid-semester Examination Jan-May 2025
BT307 Biological Data Analysis

Marks: 40

Instructions: 1) Preferably write the answers following the order of questions. 2) You MUST SHOW ALL the relevant steps of your calculation/derivation. 3) Clearly MARK the FINAL answer. 4) Use standard mathematical notations/symbols only. 5) Marks will be deducted for irrelevant calculations/derivations. 6) No marks for a partially correct answer.

Section A (each question has two marks)

Q1. A botanist is studying the variation in leaf length within a particular species. Five leaves from different plants of the same species are measured (in cm), and the results are 16, 8, 10, 14, 12. **Calculate** the sample variance from this data.

Q2. A clinical researcher is investigating whether three different antihypertensive drugs produce different reductions in systolic blood pressure among hypertensive patients. The drugs are labelled as Drug A, Drug B, and Drug C. The researcher conducted a one-way ANOVA to compare the mean reductions in pressure and obtained the following results: F-value: 5.12, Degrees of Freedom: Between groups = 2, Within groups = 21, p-value: 0.015. The level of significance $\alpha = 0.05$.

State the null hypothesis of this test and then **write** the conclusion drawn from this test.

Q3. **What** is the use of the Shapiro-Wilk test, and **what** is its Null Hypothesis?

Q4. X and Y are two independent random variables: $X \sim N(a,b)$, and $Y \sim N(p,q)$. **What** is the distribution of Z if $Z = X + Y$?

Q5. In an experiment, we have four groups. We performed all possible pairwise t-tests with a significance level $\alpha = 0.01$. **What** is the probability of having at least one type-I error in these repeated pairwise tests?

Section B (each question has six marks)

Q6. We are investigating whether the expression level of a particular oncogene differs between malignant and benign tumour tissues. The expression levels (in arbitrary units) were measured in multiple independent samples in these two groups, which are given in the table. We performed an unpaired two-sample t-test on this data. **Calculate** the t-value for this test. Assume that both populations have normal distributions and their population variances are equal.

Malignant Tumor	10	12	13	13
Benign Tumor	8	9	11	10

Q7. Researchers conducted a study to investigate the association between the presence of a specific liver enzyme marker and the severity of chronic liver disease. The study included 200 patients, and the results were summarized in the following contingency table:

	Marker present	Marker absent
Severe disease	40	10
Mild/No Disease	20	130

What is the probability that a person tests positive for the marker, given that the person has severe liver disease? Also, **calculate** the joint probability that a randomly selected person has severe liver disease and tests positive for the liver enzyme marker.

Q8. In a study on bacterial growth, researchers measured the doubling time (in minutes) of a particular bacterial strain when exposed to a new antibiotic. The following doubling times were recorded for five independent cultures: 18, 22, 19, 21, 20. Assuming that the doubling times are normally distributed, **calculate** the margin of error for the mean doubling time at a 95% confidence level. Also, **calculate** the confidence interval for the mean.

Several t-values are given below. If required, you may use them.

For df = 4: $t_{0.025} = 2.77$, $t_{0.05} = 2.13$; For df = 5: $t_{0.025} = 2.57$, $t_{0.05} = 2.01$;

The same can be written as,

For df = 4: $P(t \geq 2.77) = 0.025$, $P(t \geq 2.13) = 0.05$; For df = 5: $P(t \geq 2.57) = 0.025$, $P(t \geq 2.01) = 0.05$;

Q9. A mutation in the BRCA1 gene is known to be associated with a specific type of cancer. The prevalence of this cancer is 3%. Large-scale population-based studies have shown that 30% of patients with this cancer have this mutation. On the other hand, 1% of normal people also have this mutation. If a person has the BRCA1 mutation, **what** is the probability that the person has the cancer?

Q10. The recognition site for the restriction enzyme *Eco RI* is GTTAAC. Assume that in a DNA strand, all bases appear with equal probability. **What** is the probability of having no *Eco RI* site in a 200-base-long random DNA sequence?