Statistical Modelling and Inferencing

Practice Paper Set

TOTAL MARKS: 35

General Instructions

- Attempt ALL questions.
- Statistical tables and direct values will be provided within the question itself.

Question 1.

A mobile app company wants to estimate the proportion of users who would upgrade to a premium subscription. The company needs reliable survey results for their business planning.

Part (a) The company wants to estimate the true proportion of potential upgraders with 99% confidence and a margin of error of no more than 0.05. If no prior information is available about user preferences, what sample size should be used? **[5 marks]**

Part (b) After conducting the full survey, the company found that 156 out of 520 users indicated they would upgrade to premium. Construct a 99% confidence interval for the true proportion of users who would upgrade. Interpret this confidence interval in the context of the business decision the company needs to make. **[4 marks]**

Question 2.

A pharmaceutical researcher is investigating the effects of two factors on patient recovery time (in days): Drug Type (Standard vs. New) and Dosage Level (Low vs. High). The researcher measures recovery times and obtains the following results:

Mean Recovery Times (days):

Drug Type	**Low Dosage**	**High Dosage**	**Overall Mean**
Standard	14	10	12.0
New	8	12	10.0
Overall Mean	**11.0**	**11.0**	**11.0**

Two-Way ANOVA Results:

Source	**SS**	**df**	**MS**	**F-stat**	**P-value**
Drug Type	16.0	1	16.0	3.56	0.082
Dosage Level	0.0	1	0.0	0.00	1.000
Interaction (Drug× Dosage)	64.0	1	64.0	14.22	0.003
Error	54.0	12	4.5		
Total	134.0	15			

Part (a) Looking at the mean recovery times table, describe the pattern you observe. How do recovery times change with drug type and dosage level? Using $\alpha = 0.05$, interpret the ANOVA results for the main effects (Drug Type and Dosage Level) and the interaction effect. Which effects are statistically significant? **[4 marks]**

Part (b) Explain what the significant interaction effect means in practical terms for this pharmaceutical study. The overall mean for Dosage Level is exactly the same (11.0 for both Low and High), yet we see very different recovery times within each drug type. Explain this paradox and discuss how the researcher should interpret these findings when making recommendations about drug prescription. **[5 marks]**

Question 3.

A retail analytics team wants to predict monthly revenue (in lakhs Rs.) based on marketing spend (in lakhs Rs.) and store footfall (in thousands). Using data from 18 months, the following regression model was fitted:

Revenue = $\boldsymbol{\beta_0 + \beta_1(Marketing) + \beta_2(Footfall) + \varepsilon}$

Regression Output:

Coefficient	**Estimate**	**Std. Error**	**t-statistic**	**P-value**
Intercept (β_0)	8.2	2.5	3.28	0.005
Marketing (β_1)	2.4	0.6	4.00	0.001
Footfall (β_2)	1.8	0.9	2.00	0.064

Additional Information: $R^2 = 0.78$, Adjusted $R^2 = 0.75$, Overall F-test p-value = 0.0002

Part (a) At $\alpha = 0.05$ significance level, examine whether the overall regression model is significant. Then, analyze the individual t-tests for each predictor variable. Are both Marketing spend and Footfall significant predictors of Revenue? The manager suggests removing Footfall to simplify the model. Based on the statistical output, would you recommend this? Justify your answer. **[4 marks]**

Part (b) If the revenue data shows a clear upward trend with quarterly seasonal patterns, explain why Holt-Winters exponential smoothing would be more appropriate than simple exponential smoothing for forecasting future revenue. Describe what specific components Holt-Winters captures that simple exponential smoothing cannot. Also, explain one advantage and one disadvantage of using a 3-month moving average compared to exponential smoothing methods. **[5 marks]**

Question 4.

A technology company conducted a product trial where 30 users tested a new software feature, and 21 of them found it useful and continued using it.

Part (a) Using Maximum Likelihood Estimation (MLE), derive the estimate for the probability (p) that a user will find the feature useful. Show your complete derivation including: (i) the likelihood function, (ii) the log-likelihood function, and (iii) the derivative and solution. Interpret what this estimate tells the company about the feature's potential success. **[4 marks]**

Part (b) The company also wants to predict whether a user will adopt the feature based on their age. Historical data shows the following pattern:

Age (years)	22	25	28	32	38	42	48	52	58	62
Adopted	1	1	1	1	1	0	0	0	0	0

(1 = adopted, 0 = did not adopt)

After fitting a logistic regression model: **log(odds) = 3.2 − 0.08 × Age**

Calculate the probability that a 35-year-old user will adopt the feature using the logistic function. Then explain why logistic regression is more appropriate than linear regression for this problem, discussing at least two specific issues that would arise if ordinary linear regression were used instead. **[4 marks]**