LOYOLA COLLEGE (AUTONOMOUS), CHENNAI – 600 034

M.Sc. DEGREE EXAMINATION – **STATISTICS**

SECOND SEMESTER – **APRIL 2014**

ST 2817 - CATEGORICAL DATA ANALYSIS

Date : 05/04/2014 Dept. No. Max. : 100 Marks

Time : 09:00-12:00

SECTION – A

Answer ALL the following questions (10 x 2 = 20 marks)

1. Distinguish between 'interval scale' and 'ratio scale' of measurements.
2. What is 'overdispersion'? Give a situation where it occurs.
3. Illustrate with a numerical example how two values of 'Odds Ratio' represent the same strength of association in opposite directions.
4. Give the sample version of 'Relative Risk' (explaining the notations) and state its properties.
5. Define 'Sensitivity' and 'Specificity' in predictive diagnostics.
6. State any two properties of Yule's Q coefficient.
7. State the amended estimator of odds ratio when there are zero cell frequencies.
8. Define 'Deviance' for a Generalized Linear Model (GLM).
9. Give the motivation for Probit models..
10. Explain 'Baseline Category Logits' for nominal response variables.

SECTION – B

Answer any FIVE questions (5 x 8 = 40 marks)

11. Suppose that $P(Y_i = 1 \mid \pi_i) = \pi_i$, $i = 1,2\ldots,n$, where Y_i are Bernoulli and π_i are independent from a distribution with p.d.f. g(.). Explain why Y has Binomial distribution unconditionally, but not conditionally on the $\{\pi_i\}$.
12. Derive the Score Test and Score Confidence Interval for the Poisson parameter.
13. Discuss with examples: Case- Control study, Cohort study and Clinical Trials.
14. The following table shows the frequencies of outcomes Y (success, failure) of treatments X (A, B) given in two clinics Z (1,2) to 300 patients. Investigate marginal and conditional independence of X & Y by computing appropriate measures. Interpret your findings:

Clinic	Treatment	Outcome Success	Outcome Failure
1	A	54	36
1	B	36	24
2	A	6	24
2	B	24	96

15. Apply the 'Delta' Method to obtain the asymptotic standard error of the log odds ratio.

[cont'd]

16. The following are the computed logit scores obtained in building a binary logit model involving 20 data points:

DV	1	1	0	1	0	0	1	1	1	0
Logit	34.18	43.08	0.87	2.52	–5.51	–8.94	22.11	26.24	5.67	5.91

score										
DV	1	1	0	1	0	0	1	1	1	0
Logit score	27.36	33.84	36.48	6.45	–4.41	8.98	18.21	38.4	40.02	4.29

Construct the 'Gains Table' and compute the KS statistic for the model.

17. The following is a cross-tabulation of performance (X) in a PG program of a university and the positions/ job(Y) they were placed. Carry out the 'Linear Trend Analysis' with one-sided alternative by scoring X as 80, 65, 55 and Y as 1,2,3.

Position / Job (Y)

College Grade (X)	Data-Entry Operator	Analyst	Team-Head
Distinction	18	30	36
I Class	42	57	30
II Class	73	48	12

18. Develop the 'continuation ratio logit' for an ordinal response variable. Present the flow-chart for classification of response-category membership.

SECTION – C

Answer any TWO questions (2 x 20 = 40 marks)

19. (a) Discuss the Fisher-Rao Score test for real parameter case and multi-parameter case in a general setting.

(b) A sample of 300 washing machines of a particular brand were classified according to whether they suffered a primary mechanical failure within 90 days of installation. Machines that failed were also classified according to whether they got a secondary failure within four weeks after the first failure. Machines that did not suffer the mechanical failure cannot suffer the secondary failure. That is, no observations can fall in the category for "No" primary failure &"Yes" secondary failure (a *structural zero*).

The data are summarized below:

Secondary Failure

Primary Failure	Yes	No
Yes	58	125
No	---	117

[cont'd]

The objective was to test whether the probability of primary failure was the same as the conditional probability of secondary failure given that the machine suffered the primary failure. With appropriate parametrization, formulate the hypothesis and carry out the test. (4 + 16)

20. The following table contains the data from a survey conducted to assess the opinion of people belonging to different economic strata to a proposal of the city municipal body to build a recreation facility in a residential neighbourhood:

Economic Stratum	Support to Proposal	Neutral to Proposal	Opposed to Proposal
Low Income	81	33	112
Middle Income	105	60	48
High Income	54	30	17

(a) Compute 'Goodman and Kruskal' Gamma coefficient and interpret the result.
(b) Test the hypothesis of independence between economic strata and opinion by the Likelihood Ratio G^2 Statistic. (8 + 12)

21. (a) I x J tables, show that the odds ratio for rows a & b (a < b) and columns c & d (c < d) can be obtained from '*local odds ratios*'.
(b) Develop the Likelihood equations for a GLM and obtain an expression for the asymptotic var-cov matrix of the MLEs. (8 + 12)

22. (a) Develop the adjacent-categories logits for an ordinal response variable. Explain how the probabilities for the different response categories are estimated.
(b) In a study on academic performance with response categories 1 – Outstanding, 2 – Good, 3 – Average, 4 – Poor, the predictor variables were taken to be the following: X_1 – No. of Siblings, X_2 – 1 if High Income group, 0 otherwise.

The proportional odds model fitted gave the following cumulativelogits:

$$\text{Log } \frac{p_1}{p_{234}} = -0.842 - 0.55 * X_1 + 1.794 * X_2$$

$$\text{Log } \frac{p_{12}}{p_{34}} = 1.568 - 0.55 * X_1 + 1.794 * X_2$$

$$\text{Log } \frac{p_{123}}{p_4} = 2.041 - 0.55 * X_1 + 1.794 * X_2$$

Compute the probabilities for the performance levels for the following categories of persons: (i) High Income group with $X_1 = 1$; (ii) Middle income group with $X_1 = 3$; (iii) Low income group with $X_1 = 6$. (8 +12)