LOYOLA COLLEGE (AUTONOMOUS), CHENNAI – 600 034

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

SECOND SEMESTER – **APRIL 2022**

PST 2504 – CATEGORICAL DATA ANALYSIS

Date: 22-06-2022 Dept. No. □ Max. : 100 Marks
Time: 09:00 AM - 12:00 NOON

SECTION – A

Answer ALL the following Questions **(10 x 2 = 20 Marks)**

1. Distinguish between nominal and ordinal variables and give an example for each.
2. Define 'Information Matrix' in the context of multiparameter estimation.
3. Explain 'Poisson sampling scheme' with an example.
4. Under usual notations, state the mean and variance functions for the random component of a GLM under the 'exponential dispersion family' setting.
5. Outline the Newton-Raphson method of solving the likelihood equations of a GLM.
6. Show that for logit regression with a single binary predictor variable, the log odds ratio is same as the effect parameter.
7. Define homogeneous association in the context of 2x2xK contingency tables.
8. Define 'Continuation Ratio Logits' for ordinal responses.
9. Define a 'concordant pair' in the study of association of ordinal variables.
10. Write a note on log-log model.

SECTION – B

Answer any FIVE Questions **(5 x 8 = 40 Marks)**

11. Let Y_i be Bernoulli r.v.'s with $P(Y_i = 1) = \pi$, i = 1,2,…,n and let $Y = \sum_{i=1}^{n} Y_i$. Suppose that { Y_i } have pairwise correlation $\rho > 0$. Show that Y has overdispersion relative to $B(n, \pi)$ If π is a r.v. on (0 ,1) with mean ρ and positive variance and $P(Y_i = 1 \mid \pi,) = \pi$, show that Y has over-dispersion relative to $B(n, \rho)$.

12. Define 'Gini Concentration Index' and hence 'τ', the Concentration Coefficient. Compute 'τ' to analyse how strongly the 'opinion on recreation facility' is related to 'experience of employees' using the following data from a survey conducted among employees of a large company:

		Opinion on Recreation Facilities		
		Unnecessary	**Neutral**	**Welcome**
Employee Experience Level	**Freshers (< 1 year)**	117	292	745
	Medium Experience	624	636	758
	Senior Employees	597	208	204

13. With the data in Q. No. (12), identify which opinion is predominantly prevalent in each of the three employee groups by carrying out the Chi-square Residual Analysis [Pearson X^2 Statistic is not needed]

14. Describe the components of a GLM explaining the notations used. Identify these components for a model with a count response variable.

15. Derive the Likelihood Equations for a GLM and obtain an expression for the asymptotic covariance matrix of the MLEs.

16. A binary logit model was built using 8 records and the following observations on the dependent variable and the corresponding probability scores are reported below:

DV	1	0	0	0	1	0	1	1
Prob score	0.647	0. 543	0.192	0. 135	0.623	0. 566	0.657	0. 478

Compute the Kolmogorov-Smirnov Statistic for the model and the optimal cut-point for prediction.

17. Explain Adjacent-Category Logits and bring out the relationship with Baseline-Category Logits. Explain the method of estimating the probabilities for the outcome categories from an 'Adjacent-Category Logit Model'.

18. For the data in Q.No. (12), compute the appropriate measure of association taking into account the ordinality in the measurement scales and interpret the result.

SECTION – C

Answer any TWO Questions **(2 x 20 = 40 Marks)**

19. (a) Obtain the Wald Asymptotic confidence interval for log-odds ratio. Test whether X and Y are independent from the following contingency table (at 5% significance level) by constructing the Wald Asymptotic confidence interval for the log odds ratio:

X \ Y	Success (1)	Failure (0)
1	528	264
0	132	105

(b) Explain how the 'Partitioning of Chi-Square Statistic' is carried out in 2xJ contingency tables. Generalize it for IxJ tables. **(14 + 6)**

20. The financial support committee of a university processes the applications of students seeking financial assistance for the first semester and decides to grant assistance or reject the application. An applicant who fails to get assistance for the first semester may reapply for assistance in the second semester. The committee reviews the application and may grant assistance for the second semester or deny it completely. Students who received financial assistance in the first semester continue to get it for the second semester and do not reapply for assistance. Decisions on a sample of 850 applications for assistance over the last few years are reported below:

		Decision in second semester	
		Rejected	**Granted**
Decision in first semester	Rejected	196	390
	Granted	------	264

With appropriate parametrization, form the likelihood equation and derive the MLE of the probability of rejection in 1st semester. Test using Pearson's X^2 statistic whether the probability for rejection in 1st semester is same as the conditional probability of rejection in second semester given that there was a rejection decision in the first semester. Interpret your findings.

21. (a) Explain Case-Control studies, Cohort studies and Clinical Trials with examples.

(b) At the end of a 3-year study on a cohort of 3092 graduated students of a private university, their employment status are reported below

Gender	Major	Status	
		Unemployed	Employed
Males	Engineering	220	1773
	Others	40	301
Females	Engineering	8	142
	Others	34	574

There is a criticism that engineering graduates are suffering more than others due to unemployment and the placement cell of the university wanted to associate unemployment status to engineering majors versus other majors using difference of proportions, relative risk and odds ratio, but ignored the gender information. Interpret the results obtained by the placement cell. If you carry out the gender level analysis with the same measures, what would be your interpretation? Would you still agree that engineering graduates suffer more unemployment problem than other graduates? **(8 +12)**

22. (a) Define Adjacent-Category Logits for ordinal responses and bring out the relationship with baseline logits. Explain the estimation of probabilities of membership of an individual to the response categories.

(b) A multinomial logit model is built to relate preferred tourist spot (Temples / Recreation Parks / Monuments / Hill Stations) to age and employment sector (IT / Communication / Service Sector) of employees of a large organization. 'Hill Station' is taken as the baseline category and the coefficient of the logit equations obtained are summarized in the following table [p_T , p_R , p_M , p_H are the probabilities for preferring Temples, Recreation Parks, Monuments and Hill Stations respectively]

Logit	Intercept	IT	Communication	Age
Log[p_T / p_H]	–1.95	–1.83	0.92	0.272
Log[p_R / p_H]	1.56	–0.45	–0.63	2.06
Log[p_M / p_H]	–0.82	0.96	0.74	2.18

With the above model output, find the probabilities for preference of each of the four tourism spots by (i) a 30 year old IT employee, (ii) a 25 year old Service Sector employee. **(8 + 12)**

@@@@@@@