MERU UNIVERSITY OF SCIENCE AND TECHNOLOGY

P.O. Box 972-60200 – Meru-Kenya.
Tel: +254(0) 799 529 958, +254(0) 799 529 959, +254 (0)712 524 293
Website: www.must.ac.ke Email: info@mucst.ac.ke

UNIVERSITY EXAMINATIONS 2023/2024

FOURTH YEAR SECOND SEMESTER EXAMINATION FOR DEGREE OF BACHELOR OF SCIENCE IN STATISTICS

SMS 3458: CATEGORICAL DATA ANALYSIS

DATE: APRIL 2023 **TIME: 2 HOURS**

INSTRUCTIONS: Answer Question ONE and any other TWO questions.

QUESTION ONE (30 MARKS)

a) For a 2 x 2 table define Odds Ratio (OR) and explain how it is intelpreted. [2 marks]

ii) Using the Delta method, show that the estimated variance of the Log (RR) is given by:

$$\text{Var [log (OR)]} = \frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)} + \frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)}.$$ [4 marks]

b) A coin is flipped there times. Let Y be the number of heads obtained, when the probability of a head for a flip equals π.

i) Assuming π= 0.5 specify the probabilities for the possible values for Y and find the mean and variance of Y. [2 marks]

ii) Suppose you observe $Y = 1$ and do not know Tr. Calculate the likelihood function [2 marks]

iii) Show that the Maximum Likelihood estimate of π is $\hat{\pi} = \frac{1}{3}$ [2 marks]

c) A survey was conducted to determine whether the true proportion of shoppers favouring Brand A over Brand B is the same in two cities. Each shopper was asked whether he or she favours Brand A or Brand B. The data can be summarized in tabular form:

	Number favouring Brand A	Number favouring Brand B
Kisumu	174	93
Nakuru	196	124

(i) Construct a 90% confidence interval for the corresponding difference in proportions and interpret [4 marks]

(ii) Compute the corresponding true odds ratio and relative risk and interpret. [4 marks]

(iii) Use the x^2 test to test independence. [3 marks]

d) For the logistic regression model that has linear form for the logit of the success probability $\pi(x)$ when the random variable X takes particular value x;

$$Logit(\pi(x)) = \alpha + \beta x$$ [3marks]

Show that the relationship between $\pi(x)$ and the x can be described by the logistic function $\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$

e) The data shown in the following table refers to the effect of passive smoking on lung cancer. It summarizes results of case-control studies from three countries among nonsmoking women married to smokers.

Country	Spouse Smoked	Cases	Controls
A	No	21	82
	Yes	73	188
B	No	5	16
	Yes	19	38
C	No	71	249
	Yes	137	363

i) Describe the associations in the partial tables. [3 marks]

ii) Does the data exhibit Simpson's Paradox? [1 mark]

QUESTION TWO (20 MARKS)

a) The table below refers to a study that investigated the relationship between smoking and myocardial infarction (MI):

Ever Smoker	MI Cases	Controls
Yes	172	173
No	90	346

Compute:

i) The large sample 95% confidence interval for the odds ratio and interpret your result. [5 marks]

ii) The large sample 95% confidence interval for the relative risk and interpret your result. [6 marks]

b) To determine whether there is really a relationship between an employee's performance in the company's training program and his or her ultimate success in the job, the company took a sample of 40 cases from its very extensive files and obtained the results shown in the following table:

		Performance in training program		
Success in Employee's rating		Below average	Average	Above average
	Poor	23	60	29
	Average	28	79	60
	Very good	9	49	63

Use the 0.01 level of significance to test the null hypothesis that performance in the training program and success in the job are independent [9 marks]

QUESTION THREE (20 MARKS)

a) Let Y follow a binomial distribution wit $n = 20$ and the success probability it, and we observe Y = 0

i) Apply the Wald test for $H_0: \pi = 0.5$. Compute the corresponding 95% confidence interval. [4 marks]

ii) Apply the score test for $H_0: \pi = 0.5$. Compute the corresponding 95% confidence interval. [4 marks]

b) For all trials in Florida involving homicides between 1976 and 1987, M. Radelet and G. Pierce (*Florida Law Review*, 43,: 1-34 (1991)) reported the following results: The death penalty was given in 227 out of 4645 cases in which a white killed a white, in

92 out of 731 cases in which a black killed a white, in 9 out of 264 cases in which a white killed a black, and in 36 out of 4428 cases in which a black killed a black.

i) Exhibit the data as a three-way contingency table. [2 marks]

ii) Construct the partial tables needed to study the conditional association between defendant's race and the death penalty verdict. Compute and interpret the sample conditional odds ratios. [6 marks]

iii) Compute and interpret the sample marginal odds ratio between defendant's race and the death penalty verdict. [2 marks]

iv) Do these data exhibit Simpson's paradox? Explain. [2 marks]

QUESTION FOUR (20 MARKS)

a) The table below contains results of a study that assessed factors associated with women's attitudes toward mammography. The columns refer to their response to the question, "How likely is it that a mammogram could find a new case of breast cancer?"

Mammography Experience	**Detection of Breast Cancer**	
	Very Likely	**Not Likely**
Never	16	4
Over one year ago	12	1

Use Fisher's exact test to test:

i. $H_0 : \theta = 1$ against. $H_1 : \theta > 1$ Interpret the results (5marks)

ii. $H_0 : \theta = 1$ against $H_1 : \theta < 1$ Interpret the result (5marks)

iii. Obtain and interpret a two-sided exact P-value (3marks)
(test at $\alpha = 5\%$ level of significance)

b) The data below refers a sample of subjects randomly selected for an Italian study on the relation between income and whether one possesses a travel credit cards (such as

American Express or Diners Club). At each level of annual income in millions of lira, the table indicates the number of subjects sampled and the number of them possessing at least one travel credit card.

Income	Number of Cases	Credit Cards	Income	Number of Cases	Credit Cards
24	1	0	48	1	0
27	1	o	49	1	O
28	5	2	50	10	2
29	3	0	52	1	
30	9	1	59	1	0
31	5	1	60	5	2
32	8	O	65	6	6
33	1	o	68	3	3
34	7	1	70	5	3
35	1	1	79	1	0
38	3	1	80	1	0
39	2	o	84	1	0
40	5	O	94	1	0
41	2	0	120	6	6
42	2	0	130	1	1
45	1	1			

Answer the following questions **(see the SAS output below):**

i) Using income as the predictor variable and the number of credit cards as the response, fit a Poisson loglinear model. [1 mark]

ii) Estimate the mean number of credit cards one possesses for a person earning an income of 75 million lira. [2 marks]

iii) Conduct a Wald test of hypothesis that the mean number of credit cards is independent of income; State the alternatives, decision rule, and conclusion. [2 marks]

iv) Test goodness of fit of the model; State the alternatives, decision rule, and conclusion. (Use α= 0.05 level of significance). [2 marks]

QUESTION FIVE (20 MARKS)

a) What are the Deviance and Pearson Chi-Square criteria for determining whether the model fits the data adequately? [5 marks]

b) A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications, to get a flu shot in time for protection against an expected flu epidemic. In a pilot follow-up study, 50 clients were randomly selected and asked whether they actually received a flu shot. In addition, data were collected on their age (X_1) and their health awareness. The latter data were combined into a health awareness index (X_2) , for which higher values indicate greater awareness. A client who received a flu shot was coded Y=1, and a client who did not receive a flu shot was coded Y= 0.

Answer the following questions (see the SAS output below):

Multiple logistic regression model $Logit(\pi(x)) = \alpha + \beta_1 X_1 + \beta_2 X_2$ with two predictor variables is assumed to be appropriate.

i) Fit the multiple logistic regression model $Logit(\pi(x) = \alpha + \beta_1 X_1 + \beta_2 X_2$
Interpret the model fit. [3 marks]

ii) Test the hypotheses $H_0 : \beta_1 = 0$ *versus* $H_1 : \beta_1 \neq 0$ and $H_0 : \beta_2 = 0$ *Versus* $H_1 : \beta_2 \neq 0$ [4marks]

iii) Obtain $\exp(\beta_1)$ and interpret this number [2marks]

iv) Obtain $\exp(\beta_2)$ and interpret this number [2marks]

v) What is the estimated probability that clients aged 55 with a health awareness index of 60 will receive a flu shot? [2 marks]

vi) Obtain the model deviance and test the goodness of fit of logistic regression model obtained in part (a) above. Use α = 0.01 level of significance.[2 marks]