50

(University of Choice)

MASINDE MULIRO UNIVERSITY OF SCIENCE AND TECHNOLOGY (MMUST)

MAIN CAMPUS

UNIVERSITY EXAMINATIONS
MAIN EXAM

2022/2023 **ACADEMIC YEAR**

SECOND YEAR SECOND SEMESTER **EXAMINATION**
FOR THE DEGREE OF BACHELOR OF SCIENCE IN EPIDEMIOLOGY AND BIOSTATISTICS (BSc EPIMED)

COURSE CODE: HEM 225
COURSE TITLE: CATEGORICAL DATA ANALYSIS

DATE: 14/04/2023 **TIME: 11.00-1.00 PM**

INSTRUCTIONS TO CANDIDATES:

Answer all Questions from section A and any other two questions from section B
TIME: 2 Hours

MMUST observes ZERO tolerance to examination cheating

Paper Consists of 5Printed Pages. Please Turn Over

SECTION A (40 MKS) COMPULSORY

1. Distinguish between the following terms; (6 marks)

i. Odds ratio and relative risk

ii. Nominal scale of measurement and ordinal scale of measurement

iii. Discrete data and continuous data

2. State and describe two types of distributions used in for Categorical data analysis. (2 marks)

3. Given a random sample of size n, from a population whose pdf is;

$$f(x,\alpha,\beta)=\begin{cases}\dfrac{1}{\sqrt{2\pi\beta}}e^{\frac{-1}{2\beta}(x-\alpha)^2} & -\infty<x<\infty \quad , \quad \beta>0\\ 0, elsewhere\end{cases}$$

Obtain mle of α and β (6 marks)

4. A new medicine is tested in an experiment involving 40 patients. During the experiment, the medicine is given to 40 randomly chosen patients, and the remaining 20 patients are given a Placebo treatment. After the treatment, it is seen which patients are still ill. The result was as follows;

	fit	ill
Medicine	8	12
placebo	2	18

Investigate whether the medicine has had a significant effect. (9 marks)

5. Explain any three advantages of fitting Log linear models to categorical data. (3 marks)

6. A hypothetical cohort study in which 5000 women who used oral contraceptives and the same number who did not were followed for 10 years. The number of deaths due to myocardial infarction (Heart disease) in e ach group was recorded. 200 oral contraceptive users were lost during the follow up period due to migration and other causes.

	Death from HD	
OC Use	**Yes**	**No**
Yes	7	4793
No	2	4823

a. Estimate the risk of death from HD for women who use the OC. (2 marks)

b. Estimate the risk of death from HD for non OC users. (2 marks)

c. Estimate the relative risk and the 95% confidence interval for the relative risk. Interpret your results. (4 marks)

d. Calculate the Chi-square statistic, test the hypothesis and interpret your results. Compare with your result in (c). (6 marks)

SECTION B (answer any two questions)

QUESTION ONE (15 MKS)

a. Outline the steps used in carrying out Fishers test for independence in a 2x2 contingency table. (5 marks)

b. The following results were obtained in a study to identify whether job satisfaction was associated with income.

		Job satisfaction	
		satisfied	Not satisfied
income	Low	3	5
	High	10	7

Is there an association between job satisfaction and income at α =0.01. (10 marks)

QUESTION TWO (15 MKS)

The dataset (training) is a collection of data about some of the passengers (889 to be precise), an d the goal of the competition is to predict the survival (either 1 if the passenger survived or 0 if they did not) based on some features such as the class of service, the sex, the age etc.

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
1	0	3	male	22	1	0	7.25	S
2	1	1	female	38	1	0	71.3	C
3	1	3	female	26	0	0	7.92	S
4	1	1	female	35	1	0	53.1	S
5	0	3	male	35	0	0	8.05	S
6	0	3	male	NA	0	0	8.46	Q
……..								

VARIABLE DESCRIPTIONS:

survival	**Survival**	**(0 = No; 1 = Yes)**
pclass	**Passenger Class**	**(1 = 1st; 2 = 2nd; 3 = 3rd)**
sex	**Sex**	
age	**Age**	
sibsp	**Number of Siblings/Spouses Aboard**	
parch	**Number of Parents/Children Aboard**	

fare	**Passenger Fare**
embarked	**Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)**

The following output was obtained after logistic regression was done to the data above.

```
Call:
glm(formula = Survived ~ ., family = binomial(link = "logit"),
   data = train)
Deviance Residuals:
   Min          1Q           Median       3Q           Max
-2.6064      -0.5954      -0.4254      0.6220       2.4165

Coefficients:
             Estimate     Std. Error   z value  Pr(>|z|)
(Intercept)  5.137627     0.594998     8.635    < 2e-16 ***
Pclass       -1.087156    0.151168     -7.192   6.40e-13 ***
Sexmale      -2.756819    0.212026     -13.002  < 2e-16 ***
Age          -0.037267    0.008195     -4.547   5.43e-06 ***
SibSp        -0.292920    0.114642     -2.555   0.0106 *
Parch        -0.116576    0.128127     -0.910   0.3629
Fare         0.001528     0.002353     0.649    0.5160
EmbarkedQ    -0.002656    0.400882     -0.007   0.9947
EmbarkedS    -0.318786    0.252960     -1.260   0.2076
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1065.39  on 799  degrees of freedom
Residual deviance:  709.39  on 791  degrees of freedom
AIC: 727.39
Number of Fisher Scoring iterations: 5

Analysis of Deviance Table
Model: binomial, link: logit
Response: Survived
Terms added sequentially (first to last)

             Df      Dev      Res. Df  Resid. Dev        Pr(>Chi)
NULL                          799      1065.39
Pclass       1       83.607   798      981.79            < 2.2e-16 ***
Sex          1       240.014  797      741.77            < 2.2e-16 ***
```

```
Age          1      17.495    796      724.28          2.881e-05 ***
SibSp        1      10.842    795      713.43          0.000992 ***
Parch        1      0.863     794      712.57          0.352873
Fare         1      0.994     793      711.58          0.318717
Embarked     2      2.187     791      709.39          0.334990
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a. What is the use of residual deviance and AIC in the Logistic regression model output above. (2 marks)

b. Interpret the results above (7 marks)

c. How can the model above be improved? (3 marks)

d. Compute the probability of survival for row 1, row 3 and row 6. (3 marks)

QUESTION THREE (15 MKS)

a. Given the data below, compute the main effect parameters in the model; Using simple algebra show that ; $\log E_{ij} = U + U_{1(i)} + U_{2(j)}$ (15 marks)