MAASAI MARA UNIVERSITY

UNIVERSITY EXAMINATIONS 2018/2019

THIRD YEAR SECOND SEMISTER EXAMINATION

SCHOOL OF SCIENCE AND INFORMATION SCIENCE

DEPARTMENT OF MATHEMATICS AND PHYSICAL SCIENCES

BACHELOR OF SCIENCE IN APPLIED STATISTICS WITH COMPUTING

COURSE CODE: STA 3221

COURSE TITLE: STATISTICAL MODELLING

INSTRUCTIONS:

ANSWER QUESTION ONE AND ANY OTHER TWO QUESTIONS

SECTION ONE (30 MARKS)

a) i) State four principle steps followed when using monte-Carlo method in simulations of data (4Marks)

ii) Assuming a distribution f(x) is to be sampled using markov chain Monte Carlo-Metropolis Algorithm approach. What are the five steps that must be used (5marks)

b) Differentiate between stochastic and deterministic models citing examples in each (4marks)

c) Discuss the following modelling approaches : (6marks)

i) Fixed effect models, random effect models and mixed effects model

ii) Linear and non-linear models

iii) Multivariate models and univariate modelling

d) In building a simulation model for the petrol filling station with a single pump served by a single service man. Assuming the arrival of cars and service time are random variables. Identify the following as used in simulation (6 Marks)

i) States

ii) Events

iii) Entities

iv) Queue

v) Random realizations

vi) Distributions

d) Write the probability density function (P.D.F.) for the GLM using Gamma distribution and show the distribution is a member of the exponential family. Show the expected value and variance of the distribution (5marks)

QUESTION TWO (20 MARKS)

A survey was conducted by researchers to collect data on demographic characteristics and attitudes of residents. In 2007 the survey had two attitude items measured on a 5-point Likert scale.

Item 1: A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.

Item 2: Working women should have paid maternity leave.

Responses to these items are tabulated below;

	Item2					
Item 1	Strongly Agree	Agree	Neither	Disagree	Strongly Disagree	
Strongly Agree	90	96	22	17	2	227
Agree	102	199	48	30	5	384
Neither	7	2	5	8	0	22
Disagree	42	100	20	36	7	205
Strongly Disagree	9	18	7	10	2	46
	250	415	102	101	16	884

What's the nature of the dependency between the two items and any assumptions you made?

QUESTION THREE (20 MARKS)

The following data were collected after a food poisoning outbreak. It is suspected that the potato salad, the crab salad or both were the cause. The contingency table below shows the results of a random survey of 304 diners: whether they were sick (food-poisoned) and the food that they ate.

	Potato Salad		No Potato Salad	
	Crab Salad	No Crab Salad	Crab Salad	No Crab Salad
Not Sick	80	24	31	23
I Sick	120	22	4	0

(a) What is a generalized linear model? What is the saturated model in the context of generalized linear models?

(4 Marks)

(b) A log-linear generalized linear model with a Poisson distribution was fitted to the data. The computer output below shows the analysis of deviance table for these data. Each row of the table refers to a model containing the terms given in the

left-hand column ofthat row and all the rows above it.

	Deviance	Change in Deviance
Intercept	295.253	
Sick	294.779	0.474
Potato	169.664	125.115
Crab	7 3 . 87 1	95. 79 3
potato: crab	63.196	10.676
sick:potato	6.482	56.714
sick:crab	2.743	3.739
sick:potato:crab	4.123e-10	2.743

Find a suitable model for these data and give an interpretation. What can be concluded about the likely cause of the outbreak?

(6 Marks)

(c) How is a Pearson residual defined in this model? .

(2 Marks)

(c) Calculate the Pearson residuals for your fitted model. Do they indicate an adequate fit to the data?

(4 Marks)

(a) Another researcher suggests that a logistic regression model with sickness as response would be more appropriate for these data than the log-linear model.

Describe briefly the different aims of these two approaches, and discuss whether your colleague's suggestion is a good one.

QUESTION FOUR (20 MARKS)

In surgery, it is desirable to give enough anesthetic so that patients do not move when an incision is made. It is also desirable not to use much more anaesthetic than necessary. In an experiment, patients are given different concentrations of anaesthetic. The response variable is whether or not they move at the time of incision 15 minutes after receiving the drug.

	Concentration					
	0.8	1.0.	1.2	1.4	1.6	2.5
Move	6	4	2	2	0	0
No move	1	1	4	4	4	2
Total	7	5	6	6	4	2

a) Suggest an appropriate model to explain the impact of anaesthetic on the response variable.(6 Marks)
b) Write an R program which reads these data into R data set called ana. The program should then produce a contingency table and a glm analysis(8marks)
c) From the glm analysis below, what can you conclude between concentrations of anaesthetic and movement of patients?(6maks)

(Intercept) coef.est coef.se

	-6.469	2.418
cone	5.567	2.044

n = 30, k = 2

residual deviance = 27.8, null deviance = 41.5 (difference = 13.7):

.

QUESTION FIVE (20 MARKS)

a) Use appropriate algorithm explain the use the Monte Carlo technic to prove that Ordinal Least Square (OLS) estimators of General Linear Regression (GLR) model are BLUEs

b) Highlight four instances I that will prompt a researcher to simulate data instead of real data.

MAASAI MARA UNIVERSITY

UNIVERSITY EXAMINATIONS 2018/2019

FIRST YEAR SECOND SEMISTER EXAMINATION

SCHOOL OF SCIENCE AND INFORMATION SCIENCE

DEPARTMENT OF MATHEMATICS AND PHYSICAL SCIENCES

BACHELOR OF SCIENCE IN APPLIED STATISTICS WITH COMPUTING

COURSE CODE: STA 1208

COURSE TITLE: PRINCIPLES OF SAMPLE SURVEYS

INSTRUCTIONS:

ANSWER QUESTION ONE AND ANY OTHER TWO QUESTIONS

SECTION ONE (30 MARKS)

a) Define the following terms as used in sample survey (6Marks)

i) Sample and population

ii) Sampling frame

iii) Purposive sampling and systematic sampling

iv) Parameter

b) Why do most researchers opt for using a sample in surveys as opposed to conducting a census 4marks)

c) Differentiate with examples and approaches given between probability sampling techniques from non-probability sampling? (6marks)

d) A Researcher has taken a small survey, using an SRS, for energy usage in houses. On the basis of the survey, each house is categorized as having electric heating or some other kind of heating. The January electricity consumption in kilowatt-hours for each house is recorded *(Yi)* and the results are given below: (8marks)

Type of Heating	Number of Houses	Sample Mean	Sample Variance
Electric	24	972	202,396
Nonelectric	36	463	96,721
Total	60		

From recording existing , it is known that 16,450 of the 35,000 houses have electric heating, and 18,550 have nonelectric heating.

i) Using the sample, give an estimate and its standard error of the proportion of houses with electric heating. Does your 95% CI include the true proportion?

ii) Give an estimate and its standard error of the average number of kilowatt-hours used by houses in the city. What type of estimator did you use, and why did you choose that estimator?

(a) Explain when a design may be considered as a cluster sample. What are the first-stage and second-stage units in cluster sampling of a country like Kenya? (6marks)

QUESTION TWO (20 MARKS)

A simple random sample of 1 in 20 households in a small town provided the following data about the availability of cars and the number of adults in household3.

Number of cars *(Yi)* in the household	Adults in household *(Xi)*					Total
	1	2	3	4	5	
0	58	127	9	6	0	200
1	68	140	27	4	1	240
2	4	30	5	8	3	50
3	0	3	4	2	1	10
Total	130	300	45	20	5	500

a) Obtain point estimates, and approximate 95% confidence intervals for the following given that, $\sum \bar{x}\,\bar{y} = 795$):

i. the total number of cars in the town's households,

ii. the ratio of cars per adult in the town's households,

iii. the proportion of households with 1 or more cars per adult

(b) A survey is to be conducted on the prevalence of the common diseases in a large population. For any disease that affects at least 1% of the individuals in the population, it is desired to estimate the total number of cases, with a coefficient of variation of not more than 20%. What size of a simple random sample is needed, assuming that the presence of the disease can be recognized without mistakes?

QUESTION THREE (20 MARKS)

(d) A campus population of size N = 9000 is to be surveyed by a stratified sample for the prevalence of a certain disease, based upon three strata of respective sizes Nh = 1000,3000,5000 for h 1,2,3. The costs of sampling individuals from these strata are estimated to be respectively 40, 20, and 10 USD per person. The campus health authorities believe that roughly 1% of stratum 1, 5% of stratum 2, and 12% of stratum 3 will test positive for the disease. I)

i) What is the optimal number of individuals to sample in each stratum if the total budget for data- collection in the survey is *USD20000.*

ii) Suppose that the same population were to be sampled by SRS. About how much would the SRS cost if you want to achieve the same MSE as in (a) in estimating the proportion of the population who have the disease ?

b) An opinion poll on Kenya's health concern was conducted by the Kenya National Aids Program between April 10-15, 2011. The survey reported that 89% of adults consider AIDS as the most urgent health problem of the Kenya, with a margin of error of ±3%. The result was based on telephone interviews of 872 adults.

a. What was the target population?
b. What was the sample population?
c. How was the survey was conducted?
d. How was the sample selected?

QUESTION FOUR (20 MARKS)

(e) Foresters want to estimate the average age of trees in a stand in Mau forest. Determining age is cumbersome because one needs to count the tree rings on a core taken from the tree. In general, though, the older the tree, the larger the diameter, and diameter is easy to measure. The foresters measure the diameter of all 1132 trees and find that the population mean equals 10.3. They then randomly select 20 trees for age measurement.

Tree No.	Diameter, x	Age, y	Tree No.	Diameter, x	Age, y
1	12.0	125	11	5.7	61
2	11.4	119	12	8.0	80
3	7.9	83	13	10.3	114
4	9.0	85	14	12.0	147
5	10.5	99	15	9.2	122
6	7.9	117	16	8.5	106
7	7.3	69	17	7.0	82
8	10.2	133	18	10.7	88
9	11.7	154	19	9.3	97
10	11.3	168	20	8.2	99

i) Estimate the population mean age of trees in the stand and give an approximate standard error for your estimate.

(b) An accounting firm is interested in estimating the error rate in a compliance audit it is conducting. The population contains 828 claims, and the firm audits an SRS of 85 of those claims. In each of the 85 sampled claims, 215 fields are checked for errors. One claim has errors in 4 of the 215 fields, 1 claim has three errors, 4 claims have two errors, 22 claims have one error, and the remaining 57 claims have no errors.

i. Treating the claims as parameter 's and the observations for each field as sample's, estimate the error rate for all 828 claims. Give a standard error for your estimate.

ii. Estimate (with SE) the total number of errors in the 828 claims.

QUESTION FIVE

a) Suppose we want to estimate the average number of hours of TV watched in the previous week for all adults in some county. Suppose also that the populace of this county can be grouped naturally into 3 strata (Nairobi, Kisumu, Sayepei(rural)) as summarized in the table

Statistic	Nairobi	Kisumu	Sayepe(rural)
Nh	155	62	93
nh	20	8	12
Yh	33.90	25.12	19.00
Sh	5.95	15.24	9..36
Th	5254.5	1557.4	1767.0
Ch	2	2	3

(i) Compute a 95% confidence interval for the total number of hours of TV watched in the previous week for all adults in this county.

(ii) Estimate the total sample size needed to estimate the mean hours of TV watched in this particular county to within 1 hour with 99% probability using optimal allocation (unequal and equal costs).

(b) A local radio station carries out regular polls of its listeners on items of current interest. In one such poll listeners were asked to telephone the station and just answer "yes" or "no" to the following questions.

Do you think the government of Kenya is serious in the fight against corruption?

The poll was carried out between 8 am and 9 am one morning. At 8:30 am the announcer said the percentage of "yes" vote was 63%. When the poll closed at 9 am he announced that the percentage was 52%. List two problems associated with this method of polling and suggest why each problem might cause misleading conclusion to be drawn.

MAASAI MARA UNIVERSITY

UNIVERSITY EXAMINATIONS 2018/2019

THIRD YEAR FIRST SEMISTER EXAMINATION

SCHOOL OF SCIENCE AND INFORMATION SCIENCE

DEPARTMENT OF MATHEMATICS AND PHYSICAL SCIENCES

BACHELOR OF SCIENCE IN APPLIED STATISTICS WITH COMPUTING

COURSE CODE: STA 2219

COURSE TITLE: CATEGORICAL DATA ANALYSIS

INSTRUCTIONS:

ANSWER QUESTION ONE AND ANY OTHER TWO QUESTIONS

SECTION ONE (30 MARKS)

a) Using appropriate example define the following terms as used in data analysis (5marks)

I. Nominal measure
II. P-value
III. Ordinal measure
IV. Parameter
V. Statistic

b) In the following examples, identify the response variable and the explanatory variables. (8 marks)

i) Attitude toward gun control (favor, oppose), Gender (female, male), Mother's education (high school, college).

ii) Heart disease (yes, no), Blood pressure, Cholesterol level.

iii) Race (white, Black), Religion (Catholic, Jewish, Protestant), Vote for president (Democrat, Republican, Other), Annual income.

iv) Marital status (married, single, divorced, widowed), Quality of life (excellent, good, fair, poor).

c) According to recent UN figures, the annual gun homicide rate is 62.4 per one million residents in the United States and 1.3 per one million residents in the UK. (6marks)

i) Compare the proportion of residents killed annually by guns using the (i) difference of proportions, (ii) relative risk.

ii) When both proportions are very close to 0, as here, which measure is more useful for describing the strength of association? Why?

d) Each subject in a sample of 100 men and 100 women is asked to indicate which of the following factors (one or more) are responsible for increases in teenage crime: A, the increasing gap in income between the rich and poor; B, the increase in the percentage of single-parent families; C, insufficient time spent by parents with their children. A cross classification of the responses by gender is (6marks)

Classification				
		A	B	C

Gender	Male	60	81	75
	Female	75	87	86

e) Is it valid to apply the chi-squared test of independence to this 2 × 3 table? Explain.

f) Explain how this table actually provides information needed to cross classify gender with each of three variables. Construct the contingency table relating gender to opinion about whether factor A is responsible for increases in teenage crime.

e) Based on murder rates in Kenya, a survey has reported that the probability a newborn child of eventually being a murder victim is 0.0263 for Urban males, 0.0049 for rural males, 0.0072 for rural females, and 0.0023 for white urban females.(5marks)

i) Find the conditional odds ratios between region and whether a murder victim, given gender. Interpret.

ii). If half the newborns are of each gender, for each region, find the marginal odds ratio between race and whether a murder victim.

QUESTION TWO(20 MARKS)

A doctor is investigating the effect of a woman's age on the success of an IVF (in vitro fertilisation) procedure. She has randomly selected 10 women aged under 35 and 10 women aged at least 35. From hospital records she has obtained the following data, which record the numbers of eggs obtained from the women and the numbers that were fertilized during one IVF procedure. She wants to investigate the effect of the woman's age on the probability of an egg being successfully fertilised. She calls this probability the "fertilization rate".

Women aged under 35		Women aged at least 35	
Number ofeggs	*Number of fertilised*	*Number ofeggs*	*Number of fertilised*
10	9	7	6
9	7	10	7
7	5	9	5
5	3	8	4
10	9	6	4
7	7	5	1
9	5	7	4
8	8	6	4
7	2	5	2

7	5	7	5

a) Carry out a suitable exploratory analysis to see whether the fertilization rate might depend on the woman's age.

b) Let n_j denote the number of eggs and x_i the number of fertilized eggs for the i^{th} woman. Let t_j denote the fertilization rate for the i^{th} woman.Explain why a binomial distribution may be valid to model the data.

QUESTION THREE (20 MARKS)

Discuss the following concepts as used in categorical data modeling

i) Multinomial sampling

ii) Poisson sampling

iii) Goodness of fit test

iv) Test of association

v) Relative risk and odds ratio

QUESTION FOUR (20 MARKS)

A chi-squared variate with degrees of freedom equal to *df* has representation $Z_1^2 + \cdots + Z_{df}^2$, where Z_1, . . . , Z_{df} are independent standard normal variates.

a. If Z has a standard normal distribution, what distribution does Z^2 have?

b. Show that, if Y_1 and Y_2 are independent chi-squared variates with degrees of freedom df_1 and df_2, then $Y_1 + Y_2$ has a chi-squared distribution with $df = df_1 + df_2$.

QUESTION FIVE

Table below comes from one of the studies of the link between lung cancer and smoking. The study was done in 20 hospitals patients admitted with lung cancer in the previous year were queried about their smoking behavior. For each patient admitted, researchers studied the smoking behavior of a non-cancer control patient at the same hospital of the same sex and within the same 5-year grouping

on age. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year

		Lung Cancer	
		cases	**Control**
Have smoked	**Yes**	**688**	**650**
	no	**21**	**59**
	Total	**709**	**709**

.

a. Identify the response variable and the explanatory variable.

b. Identify the type of study this was.

c. Can you use these data to compare smokers with nonsmokers in terms of the proportion who suffered lung cancer? Why or why not?

d. Summarize the association, and explain how to interpret it.

MAASAI MARA UNIVERSITY

UNIVERSITY EXAMINATIONS 2018/2019

THIRD YEAR FIRST SEMISTER EXAMINATION

SCHOOL OF SCIENCE AND INFORMATION SCIENCE

DEPARTMENT OF MATHEMATICS AND PHYSICAL SCIENCES

BACHELOR OF SCIENCE IN APPLIED STATISTICS WITH COMPUTING

COURSE CODE: STA 1209

COURSE TITLE: STATISTICAL COMPUTING AND DATA ANALYSIS

INSTRUCTIONS:

ANSWER QUESTION ONE AND ANY OTHER TWO QUESTIONS

Section one(30 marks)

i). What are the following binary values in decimal? (4marks)

a) 0000101_2

b) 0001001_2

c) 0001101_2
d) 0010101_2

ii) Differentiate between the following terms as used in computer (6marks)

a) RAM and ROM
b) Compiler and CPU
c) Input and output devices

iii)Discuss what is meant by bit and how words and character are represented in a computer(5marks)

iv)Given the following data points be (0, 2) and (2,4) use polynomial $P_1(x)$ to represent this and sketch the curve (4marks)

v)Given $f(x) = \sin(x)$, $x_0 = 0.2$, $x_1 = 0.3$. use the first-order divided difference of f(x) to approximate cos(x) (3marks)

vi) Evaluate $f(x) = e^x$, $x\epsilon[0, 1]$ and consider the error in linear interpolation to f(x) using x_0, x_1 satisfying $0 < x_0 < x1<1$ (5marks)

vii)Differentiate between linear and nonlinear equation (3marks)

QUESTION TWO(20 MARKS)

How are the following data processing concepts related:

a. Coding scheme vs. Data dictionary
b. . Data set vs. Database
c. FIat ASCII file vs. Hierarchical ASCII file
d. Editing for analysis vs. In-house editing
e. Value labels vs. Variable labels

QUESTION THREE (20 MARKS)

a) Use Newton's Method to determine x_2 for the given function and given value of x_0

i) $f(x)=x^3-7x^2+8x-3$, $x_0=5$

ii) $f(x)=x\cos(x)-x^2$, $x_0=1$

b) Using Newton's Method find the root of the given equation, accurate to six decimal places, that lies in the given interval.

i) $x^4-5x^3+9x+3=0$ in $[4,6]$

ii) $x^2+5=e^x$ in $[3,4]$

c)State the function of operating systems in a computer

QUESTION FOUR (20 MARKS)

a) Determine the Taylor Series for the given function.

1. $f(x)=\cos(4x)$ about $x=0$
2. $f(x)=x^6e^{2x}$ about $x=0$
3. $f(x)=e^{-6x}$ about $x=-4$
4. $f(x)=\ln(3+4x)$ about $x=0$

b) state fou software that can be used for data analysi

QUESTION FIVE

a) Discuss the numbers system in a computer ad using example illustrate how your convert numbers in a computer with reference to decimal number system.

b) Discuss the fundamental rules of coding in survey data processing.