# MAASAI MARA

# UNIVERSITY

## REGULAR UNIVERSITY EXAMINATIONS
## 2018/2019 ACADEMIC YEAR
## SECOND 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**

**DATE: 24TH APRIL, 2019 TIME: 0830AM – 10.30AM**

**INSTRUCTIONS:**

**ANSWER QUESTION ONE AND ANY OTHER TWO QUESTIONS**

*This paper consists of 4 printed pages. Please turn over.*

## 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 | | | | |
|---|---|---|---|---|
| Gender | | A | B | C |
| | Male | 60 | 81 | 75 |
| | Female | 75 | 87 | 86 |

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

b) 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 $Z_1, \ldots, Z_{df}$ 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.

**//END**