JARAMOGI OGINGA ODINGA UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF HEALTH SCHENCES

UNIVERSITY EXAMINATION FOR THE MASTERS IN PUBLIC HEALTH
(BIOSTATICS & EPIDEMIOLOGY)
1st YEAR SEMESTER TWO 2019/2020 ACADEMIC YEAR
KISUMU

COURSE CODE: HES 5123

COURSE TITLE: ADVANCED BIOSTATISTICS

EXAM VENUE: **STREAM**

DATE: 9/12/19 **EXAM SESSION: 2.00 – 5.00PM**

TIME: 3 HOURS

Instructions:

1. **Answer ANY 4 questions**
2. **Candidates are advised not to write on the question paper**
3. **Candidates must hand in their answer booklets to the invigilator while in the examination room**

Question 1

a. Write 8x + 4y=16 in slope-intercept form **(1Mark)**

i. How do you interpret b1 in simple linear regression? **(1Mark)**

ii. How do you interpret b1 in multiple linear regression? **(1Mark)**

iii. What is the difference between R^2 and adjusted R^2 **(2Mark)**

b. Molly earned a score of 940 on a national achievement test. The mean test score was 850 with a standard deviation of 100. What proportion of students had a higher score than Molly? (Assume that test scores are normally distributed) **(2Marks)**

c. A study was done to compare the lung capacity of coal miners to the lung capacity of farm workers. The researcher studied 200 workers of each type. Other factors that might affect lung capacity are smoking habits and exercise habits. The smoking habits of the two worker types are similar, but the coal miners generally exercise less than the farm workers (**2 Marks**)

i. Identify the outcome variable of interest?

ii. Is the outcome variable quantitative or qualitative?

iii. What is the implied population?

iv. What are the explanatory variables in this case?

d. A researcher follows 200 women who exercise regularly and 300 women who do not exercise regularly. After 30 years of follow-up, 25 of the women in the exercise group are diagnosed with osteoporosis while 30 women in the non-exercise group are diagnosed with osteoporosis.

i. Draw the 2X2 contingency table showing the disease on top and the exposure on the side. **(1 Marks)**

ii. Calculate odds ratio (**OR**) & relative risk (**RR**) of developing osteoporosis between the two groups. (Show your work.) **(3 Marks)**

e. A random sample of 20 observations produced a sample mean of $\bar{x}$ = 92.4 and s = 25.8. What is the value of the standard error of $\bar{x}$? **(2Marks)**

Question 2

a. You buy a package of 122 Smarties and 19 of them are red. What is a 95% confidence interval for the true proportion of red Smarties? **(2Marks)**

b. State four MAJOR assumptions of ANALYSIS OF VARIANCE **(2Marks)**

c. Differentiate between Student T-test and a One-Way ANOVA? **(2Marks)**

d. Define the following terms:-

i. P-value **(1Mark)**
ii. Level of significance **(1Mark)**
iii. Level of confidence **(1Mark)**

e. On average lightning kills three people each year in the UK at a rate of λ =3
What is the probability that utmost 2 persons are killed this year? **(3Marks)**

f. A significance test for comparing two means gave t=−1.97 with 10 degrees of freedom. Can you reject the null hypothesis that the μ's are equal versus the two-sided alternative at the 5% significance level? **(3Marks)**

Question 3

a. Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the treatments (cars types). Using the hypothetical data provided below, test the hypothesis that mean pressure applied to the driver's head during a crash test is equal for each types of car. Use α =5% .
State the hypothesis, calculate the appropriate test statistics and and make a conclusion **(15 Marks)**

Compact cares	Midsize cars	Full-size cars
643	469	484
655	427	456
702	525	402

Question 4

a. Are the following statements **TRUE** or **FALSE**? **(3Marks)**

A) Rejecting the ANOVA null hypothesis implies that all the means are different from one another.

B) An ANOVA test can be used if the largest standard deviation is less than twice the smallest standard deviation.

C) Multiple-comparisons methods are used only after rejecting the ANOVA null hypothesis.

D) In two-way ANOVA there are two null hypotheses.

E) For using a repeated-measures design, we require that all observations be independent.

F) In a two-way ANOVA, F statistics and P-values are used to test hypotheses about the main effects and the interaction.

b. What is the basic assumption of multilevel modelling?) **(1Mark)**

i. That the dependent variable is continuous
ii. The expected value of *Y* can be modeled by a combination of unknown parameters
iii. That the units of analysis are measured over time
iv. A unit at the lowest level is nested into a higher level unit(s)

c. What is a level? **(1Mark)**

i. A level is a given variable chosen from your theoretical approach
ii. A level is a variable that identifies units sampled from a population
iii. A level can be any categorical variable in your dataset
iv. A level can be any continuous variable in your dataset
v.

d. Which of the following are advantages of multilevel modeling? **(1Mark)**

i. It is the statistical method that comes closest to experiments in establishing causality
ii. It takes into account the problem of dependency among observations
iii. It allows us to model the influence of variables from all levels on our *Y*
iv. It allows us to model risk-development over time of an event taking place

e. In linear regression, failure to account for correlation among observations within clusters will: (**choose all correct answers**) **(1Mark)**

i. give biased estimates of regression coefficients
ii. give incorrect standard errors, but correct p-values
iii. give incorrect standard errors and incorrect p-values

iv. give you a real dilemma when your scientific paper is repudiated after being published

f. Do they follow **Poisson** distribution? Y/N, and give a reason **(4Marks)**

a. The number of heart attacks in Brighton each year

b. The number of planes landing at Hearthrow between 8 and 9am?

c. The number of cars getting a puncture at Thika road each year

d. The number of people in the UK flooded out of their homes in July

g. The life of a cat is normally distributed with a mean of 15 years and a standard deviation of 2 years. Find the probability of a cat living at least 10 years **(2Marks)**

Question 5

a. Explain what is meant by censored observations in survival data? **(1Mark)**

i. State the difference between right and left censoring, and illustrate by giving practical example **(2Marks)**

b. A large number of individuals were sampled randomly from a population of adults, were enrolled in a study and were followed for 30 years to assess the age at which a disease symptom first appeared

i. What are the time origin and time scale for this study? **(1Mark)**

ii. What is the failure event for this study? **(1Mark)**

iii. What are the common causes of censoring?

c. X is a normally normally distributed variable with mean μ = 30 and standard deviation σ = 4. Find **(3Marks)**

a) $P(x < 40)$

b) $P(x > 21)$

c) $P(30 < x < 35)$

d. A statement whose validity is tested on the basis of a sample is called? **(1Mark)**

a) Null Hypothesis

b) Statistical Hypothesis

c) Simple Hypothesis

d) Composite Hypothesis

e. State **3** benefits of randomization? **(3Marks)**

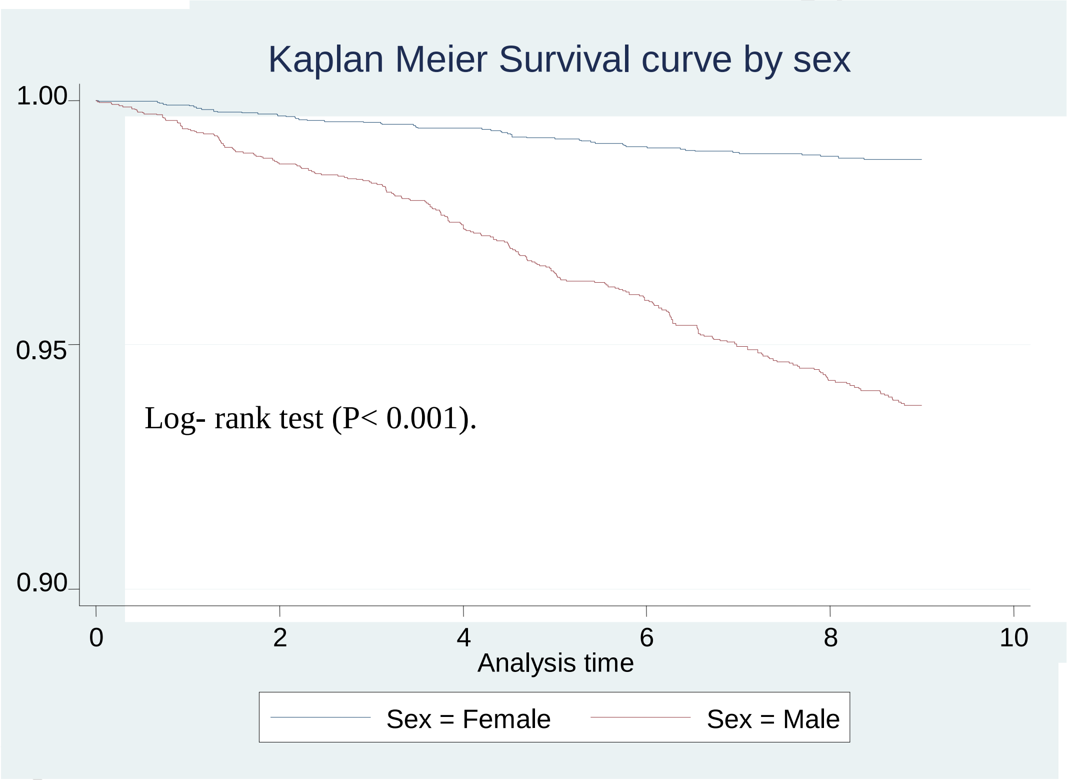
f. What is Quasi-Experimental trial? **(1Mark)**

g. If new cases of West Nile in New England are occurring at a rate of about 2 per month, then what's the probability that exactly 4 cases will occur in the next 3 months? **(2Marks)**

$$P(X = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

Questions 6

a. The Kaplan-Meier curves below shows cumulative probabilities of homicide deaths by sex over 9 years period. Interpret the graph **(4Marks)**



b. A manufacturing process produces TV. tubes with an average life m=1200 hours and **s** = 300 hours. A new process is thought to give tubes a different life. And out of a sample of 100 tubes we find that they have an average life = 1265 hours. Is the new process any different from the old process? . Test the hypothesis using both **p-value** & **critical value** method.

Show your work, including sketch of a rejection region **(6Marks)**

c. Below is an output of linear regression with a dependent variable being gallons of fuel consumed. Write a statistical model to predict gallons consumed when given insulation & temperature, and interpret your findings in terms of **R-squared & R-squared adjusted (5Maks)**

```
--------------------------------------------------------------------------------

                                   Standard         T

 Parameter            Estimate       Error       Statistic      P-Value

--------------------------------------------------------------------------------

CONSTANT               562.151      21.0931       26.6509       0.0000

Insulation            -20.0123      2.34251      -8.54313       0.0000

Temperature           -5.43658     0.336216      -16.1699       0.0000

--------------------------------------------------------------------------------
```

R-squared = 96.561 percent

R-squared (adjusted for d.f.) = 95.9879 percent