JARAMOGI OGINGA ODINGA UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF AGRICULTURAL AND FOOD SCIENCES
UNIVERSITY EXAMINATION FOR DEGREE OF DOCTOR OF PHYLOSOPHY IN FOOD SECURITY AND SUSTAINABLE AGRICULTURE

1st YEAR 1st SEMESTER 2019/2020 ACADEMIC YEAR

MAIN REGULAR

COURSE CODE: AFB 6112

COURSE TITLE: Advanced Statistics and Research Methods

EXAM VENUE: **STREAM: (PhD. In Food Security & Sustainable Agriculture)**

DATE: 3/09/19 **EXAM SESSION: 9.00 – 12.00NOON**

TIME: 3.00 HOURS

Instructions:

(i) Answer any three questions.
(ii) Candidates are advised not to write on the question paper.
(iii) Candidates must hand in their answer booklets to the invigilator while in the examination room.
(iv) Where necessary, computations and data analysis to be done with R statistical software.

Examination data set

Throughout the exam, a dataset containing demographic information and standardized test scores of past postgraduate (MSc and PhD) students will be used. Performance of 5 course units shared in common among the postgraduate students in the university were evaluated. The dataframe contains 200 rows (number of students) and 11 variables with the following characteristics.

id = identification number of the student
gender = sex of the student (male=0 or female=1)
location = Regional background of the student (Local=0 or regional=1)
ses = socio-economic status of the student (low = 0, middle = 1, high = 2)
course = Course type of the student (PhD or Msc) (Msc = 0, PhD = 1)
prog = Program type (part-time = 0, fulltime = 1)
statmethods = Statistical methods
researchmethods = Research methods
entrepreneurship = Entrepreneurship
climatechange = Climate change
statcomputing = Statistical computing

QUESTION ONE

a. A hundred and twenty chicks were subjected to a certain feed and the increase in their weight measured after one week. The increase in weight was recorded as follows:

Additional weight in grams	10-14	15-19	20-24	25-29	30-34	35-44	45-59	60-89	90-119
No of chicks	2	5	17	33	27	25	7	3	1

i. Obtain the estimates of the median and quartiles of this distribution. **(6 marks)**

ii. Comment on the skewness of the distribution. **(4 marks)**

b. Use the data provided and R to answer the following questions. Create a new variable that sum performance of the available 5 course units. Call this variable totalscore.

i) Use a boxplot to summarize the distribution of total score. Interpret the graph accordingly. **(2 marks)**

ii) Provide a summary statistic of total score by various levels of ses (socio-economic status of the students)

a. Mean **(2 marks**)

b. Sd **(2 marks)**

iii) Create a new variable honors if and only if totalscore ≥ 300

$$\text{Let } \ honors = \begin{cases} 1, \ \text{totalscore} \ \geq \ 300 \\ 0, \ \text{totalscore} \ < \ 300 \end{cases}$$

Provide Mean and Sd for each of the level of the binary variable (honors). **(2 marks)**

iv) Evaluate statistical association between honors and gender. Use chi-square test to assess the level of the association. State the hypothesis tested. **(2 marks)**

QUESTION TWO

a) The results of investigations as carried out in part (a) above were recorded as follows.

		Blocks			
		1	**2**	**3**	
Treatments	**1**	21	24	34	
	2	25	33	30	
	3	31	34	38	
	4	17	39	32	

Test the hypotheses

i. $H_{01}: t_1 = t_2 = t_3 = t_4 = 0$

ii. $H_{02}: b_1 = b_2 = b_3 = 0$ at 0.05 level of significance **(10marks)**

b. Use the data provided and R to answer the following questions. Consider the postgraduate students simulated dataset provided. Having created totalscore variable that sums performance in the 5 course units, conduct a two – way ANOVA classification involving totalscore (y) and treatments - gender and ses (socio-economic variable). Based on the model, answer the following questions:-

i) Obtain mean and summary statistics by gender and ses grouping variables **(2 marks)**

ii) Provide an interaction plot using summary statistics by the grouping variables. **(2 marks)**

iii) Fit a linear model and conduct ANOVA. Provide summary output and interpret the results. **(2 marks)**

iv) Provide an ANOVA table of the fit in (iii) and interpret the results. **(2 marks)**

v) Test whether or not the interactions effect is significant. **(2 marks)**

QUESTION THREE

a) During the lambing season 8 ewes and the lambs they bore were weighted at the time of birth with the following results

Weight of ewe[x] in kg	44	41	43	40	41	37	38	35
Weight of lamb [y] in kg	3.5	2.8	3.2	2.7	2.9	2.5	2.8	2.6

Assuming that $\sum xy = 923.2, \sum x^2 = 12785, \sum y^2 = 66.88$, calculate the product moment correlation coefficient between X and Y. At 5% level of significance, test whether the data could have come from a population with correlation coefficient $\rho = 0$ **(10 marks)**

b. Use the data provided and R to answer the following questions. Fit a multiple linear regression (OLS) with generalized linear model (glm) function involving dependent variable totalscore (y) and independent variables (statmethods, researchmethods, entrepreneurships, climatechange and statcomputing). Use this model to answer the following questions.

i) Obtain the regression model based on the summary statistics – call this model 1 (or full model). **(2 marks)**

ii) Use forward variable selection procedure to obtain most parsimonious model to this data (the simpler the model the better) – call this model 2 (reduced model) **(2 marks)**

iii) Provide R- squared to the model in (ii). Compare and contrast coefficient of determination obtained in (ii) and that of the full model. **(2 marks)**

iv) Obtain a correlation matrix involving only the independent variables obtained after the variable selection process in (ii). **(2 marks)**

v) Fit another reduced model (model 3) which includes location variable (origin of the student) besides the variables selected in model 2. Explain whether or not inclusion of the location variable is statistically significant in the model. **(2 marks)**

QUESTION FOUR

a. Discuss any three core functions of literature review in research. **(3 marks)**

b. The ages (in months) at which 50 children were first enrolled in a preschool are listed below.

38,   40, 30, 35, 39, 40, 48, 36, 31, 36, 47, 35, 34, 43, 41, 36, 41, 43, 48, 40
32, 34, 41, 30 ,46 ,35, 40 ,30 ,46, 37, 55, 39, 33, 32, 32, 45, 42, 41, 36, 50
42, 50, 37, 39, 33, 45, 38, 46, 36, 31

i. Construct a stem and leaf display for the data. Start the lower boundary of the first class at 30 and use a class width of 5 months. **(3marks)**

ii. Construct a grouped frequency distribution for the data **(4marks)**

c. Use the data provided and R to answer the following questions. Consider creating a variable similar to the one created in question one (iii)

Let $honors = \begin{cases} 1, \text{ totalscore } \geq 300 \\ 0, \text{ totalscore } < 300 \end{cases}$

The variable honors is then considered binary response (outcome, dependant). In addition, we need to assess whether five independant variables (gender, location, ses, course and prog) of the student has statistical significance on admission to honors. To answer the question, fit a binary logistic regression model to this data – call this full model.

i) Obtain summary statistics **(2 marks)**

ii) Interpret every individual slope parameters obtained from the regression analysis. **(2 marks)**

iii) Provide the regression model including the estimated slope parameters. **(2 marks)**

iv) Give a pseudo R- squared and interpret the result. **(2 marks)**

v) Fit an alternative model only with gender, ses and prog and call this a reduced model. Compare full and reduced models statistically. **(2 marks)**

QUESTION FIVE

a. Explain the following terms as used in sample surveys:

i. Sampling unit **(1 marks)**

ii. Sampling frame **(1 marks)**

iii. Purposive sample **(1 marks)**

iv. Simple Random Sampling without Replacement. **(1 marks)**

v. Cluster sampling **(1 marks)**

b. Use the data provided and R to answer the following questions. Suppose a regression model involving totalscore (y) and independant variable (statmethod) upto second order is developed

totalscore = β_0 + β_1*statmethod + β_2*statmethod^2 + Ɛ

i) What class of regression model would this be classified? **(3 marks)**

ii) Develop an appropriate regression model involving totalscore (y) and statmethod (x) upto second order. Provide the fitted model. **(3 marks)**

iii) Explain whether or not inclusion of quadratic terms is significant in the model? **(3 marks)**

iv) Plot a scatter plot with the best fitting polynomial curves. **(3 marks)**

v) Give coefficient of determination and interpret the result. **(3 marks)**