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

FIRST YEAR FIRST SEMESTER 2020/2021 ACADEMIC YEAR

REGULAR

COURSE CODE: AFB 6112

COURSE TITLE: ADVANCED STATISTICS AND RESEARCH METHODS

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

DATE: 16/2/21 **EXAM SESSION: 9-12 NOON**

TIME: 3.00 HOURS

Instructions:

(i) Answer any three questions.
(ii) Where necessary, computations and data analysis to be done with R statistical software.

2. Candidates are advised to write on the text editor provided, or to write on a foolscap, scan and upload alongside the question

3. Candidates must ensure they submit their work by clicking "finish and submit attempt" button at the end.

QUESTION ONE (20 MARKS)

a. During a recent agricultural show, two judges were assigned the duty of awarding marks for a given category of produce exhibited at different stands. Each judge was under strict instructions to give their own honest opinion on the exhibition. The marks awarded by each of the two judges to 8 exhibitors were recorded as follows

Judge 1	22	27	24	17	20	22	16	13
Judge 2	28	23	25	14	26	17	20	15

i. Calculate the Spearman's rank correlation between the marks awarded by the judges. **(2marks)**

ii. Stating your hypotheses and using a 5% level of significance, interpret your results. **(8 marks)**

b. Create a two-column data frame. Let the first column be name "judge" and second column be "marks"

i. Give the two types of data expressed in each of the two variables **(1mark)**

ii. Provide mean and standard deviation of marks awarded by judge and corresponding R-code used. **(1mark)**

iii. Develop a scatter diagram overlay the best fitting line with R. **(1mark)**

iv. State the shape of the relationship displayed. **(1mark)**

v. Add the best fitting line to the scatter diagram and interpret. **(1mark)**

vi. Fit a regression equation of marks on the categorical variable (judge)

- Write down the regression model. **(1mark)**
- Interpret the regression parameter estimates. **(1mark)**
- State the fit statistic and interpret. **(1mark)**
- What is the relationship between R^2 and the Spearman's correlation coefficient obtained in 1(a)? **(2marks)**

QUESTION TWO (20 MARKS)

a. At JOOUST, two machines A and B are used to pack cricket biscuits. A random sample of ten packets done by each machine are picked and the masses recorded in grammes. One wishes to check which between the two machines produces masses with greater variability within. Using the data from the two machines given below, compute the coefficient of variation for each machine and offer your expert opinion on the within variability.

Machine A	196	198	198	199	200	200	201	201	202	205
Machine B	192	194	195	198	200	201	203	204	206	207

(10 marks)

i. Test equality of the two variances for machine A and machine B, provide R-code

- State the hypothesis tested. **(1 mark)**
- Test statistic. **(1 mark)**
- Conclusion. **(1 mark)**

ii. Given the test results in (i), what is the most appropriate test procedure for comparing equality of the two sample means. Justify your answer. **(2 marks)**

iii. State the hypothesis and write down the test statistic. **(2 marks)**

iv. Based on the summary R output, what is the conclusion regarding equality of the two sample means? **(1 mark)**

v. Compare and contrast the results in (iv) and the conclusions obtained in 2b. **(2 marks)**

QUESTION THREE (20 MARKS)

a) An investigator carried out an experiment where 6 replicates of four treatments were observed. The experimental field was considered homogeneous. The yield per replicate of treatments was recorded as follows.

Treatments		**Observations**					
		1	**2**	**3**	**4**	**5**	**6**
	1	7	8	15	11	9	10
	2	12	17	13	18	19	15
	3	14	18	19	17	16	18
	4	19	25	22	23	18	20

Are the treatment effects the same? If not, use LSD value 3.07 to identify the treatments which are different from the rest. **(10marks)**

b) Use R to test assumptions, interpret the results and provide R-code for each case,

i. Independence of observations. **(1 mark)**

ii. Normality of the response variable. **(1 mark)**

iii. Homogeneity of variance. **(1 mark)**

iv. Provide an ANOVA table summary and interpret the results. **(1 mark)**

v. Adopt appropriate regression procedure to test significance of the various levels of the treatment (factor). Provide R-code used. **(1 mark)**

- Write down the regression equation. **(1 mark)**
- Interpret the regression parameter estimates. **(1 mark)**
- Provide and interpret the goodness-of-fit statistic. **(1 mark)**
- Sketch the scatter plot and overlay the best fitting line. **(1 mark)**
- Compare and contrast the results and conclusion obtained in 3b(v) and 3(a**) (1**

mark)

QUESTION FOUR (20 MARKS)

a. You are provided with the following data: 30, 35, 45, 48, 52, 65, 75. Compute the measures of skewness and kurtosis hence make an opinion on normality of the data. **(10 marks)**

b. Fit a probability distribution function for each case to the data in 4(a), estimate and interpret the parameters of the equation. Use R.

i. Suppose the data in 4(a) is continuous realizations from a normally distributed random variable, estimate mean and standard deviation hence or otherwise write down the resulting expression of the probability density function. **(3 marks)**

ii. Suppose the data follows a gamma distribution, provide shape, scale parameter estimates and write down the resulting probability density function . **(3 marks)**

iii. Suppose the data in 4(a) is a realization of count outcomes, fit a Poisson distribution to the data and use chi-square test procedure to assess the goodness-of-fit.**(4 marks)**

QUESTION FIVE (20 MARKS)

a. The heights (in centimetres) at which 50 seedlings were transplanted 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

Construct a grouped frequency distribution for the data starting at 30 hence describe the distribution of the data based on a histogram and frequency polygon. **(10 marks)**

b. Use the data in 5(a) to answer the following questions

i. Obtain the summary statistics of the ungrouped variable height: minimum, 1st quartile, median, 3rd quartile, and maximum. **(1mark)**

ii. Using R, develop a table of height group and frequencies taking ranges 30-35, 35-40, 40-45, 45-50 and 50-55.
Hint: Let
height = c(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)
heightcut = cut(height, breaks=c(30,35,40,45,50), right=FALSE) **(1mark)**

iii. Add percentage of the frequency's column. Provide R-code. **(1mark)**

iv. Provide column totals. Provide R-code. **(1mark)**

v. Convert the table to a data frame. Provide R-code. **(1mark)**

vi. Interpret the distribution of data based on percentages. **(1mark)**

vii. Write down the 5 by 3 data frame/table retaining only Grouped scores, Frequency,

Percentage columns. **(1mark)**

viii. Provide a modal class height. **(1mark)**

ix. Which height group registered the highest and lowest frequencies? **(2marks)**