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

UNIVERSITY EXAMINATION FOR DEGREE OF MASTER PUBLIC HEALTH

1ST YEAR 2ND SEMESTER 2022/2023 ACADEMIC YEAR

KISUMU CAMPUS

COURSE CODE: HES 5136
COURSE TITLE: STATISTICAL METHODS
EXAM VENUE: **STREAM:**
DATE: **EXAM SESSION:**
TIME: **3.00 HOURS**

Instructions:
1. Answer **Question ONE (compulsory)** and any other THREE questions.
2. Candidates MUST not to write anything on the question paper.
3. Candidates MUST hand in their answer booklets to the invigilator while in the examination room.

SECTION A

Answer question one(Compulsary)

1. Question one (10 marks).

Sixty-five pregnant women at high risk of pregnancy-induced hypertension participated in a randomized controlled clinical trial comparing 100mg of aspirin daily and a matching placebo during the 3rd trimester of pregnancy. The observed numbers with hypertension are shown in the following table.

	Hypertension		
	yes	no	Total
Aspirin	5	29	34
Placebo	10	21	31
Group Total	15	50	65

- i. Give the **estimate** and **approximate 90% confidence interval** for the following of hypertension between aspirin and placebo treated women
  - a. Difference in risk (2 marks)
  - b. Risk ratio (3 marks)
  - c. Odds ratio (3 marks)
- ii. Suppose a new study is planned. What sample size is approximately needed in order to have a power of 90% if the risk of hypertension is 0.05 lower in aspirin treated women ($\alpha = 0.05$)? (2 marks)

SECTION B

Answer any three Questions

2. Question two (20 marks).

The table below contains an historic dataset on the height (in inches) of brother and sister, published in the second volume of Biometrika by Karl Pearson (indeed the man of the correlation coefficient) (Pearson K, Lee A. On the laws of inheritance in man. Biometrika 1902; 2:p357. In this case, we look at: - the correlation between the height of brother and si
ster; how to predict height of the brother based on height of the sister or vice versa; - the difference in mean height between brothers and sisters.

famil	Height of	Height of	differen
y	brother	sister	ce

1	71	69	2	**7**	70	65	5
2	68	64	4	**8**	73	64	9
3	66	65	1	**9**	72	66	6
4	67	63	4	**10**	65	59	6
5	70	67	3	**11**	66	62	4
6	71	62	9				

Study the SPSS output given below (the SPSS names of variables are: BROTHER, SISTER)
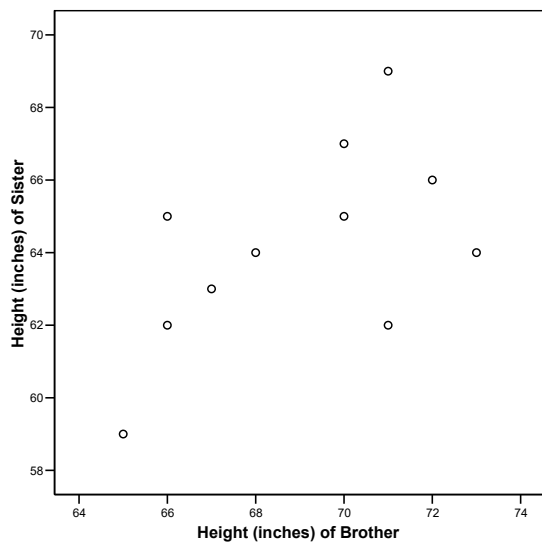
		Brother	Sister
Brother	Pearson Correlation	1	.555
	Sig. (2-tailed)	.	???
	Sum of Squares and Cross-products	74.000	41.000
	Covariance	7.400	4.100
	N	11	11
Sister	Pearson Correlation	.555	1
	Sig. (2-tailed)	???	.
	Sum of Squares and Cross-products	41.000	73.636
	Covariance	4.100	7.364
	N	11	11

Descriptive Statistics

	Mean	Std. Deviation	N
Brother	69.00	2.720	11
Sister	64.18	2.714	11

Correlations

a. Is Pearson's correlation between height of sister and brother significantly different from zero? (3 marks)
   i. Give bounds for the p-value (1mark)
   ii. Compute also an approximate 95% confidence interval (2 marks)
   iii. Comment on the appropriateness of these assumptions for these data(1mark)
b. Compute the Spearman's correlation coefficient. (3 marks)
   i. Is it statistically different from zero? (1mark)
   ii. Give the upper bound for the p-value (1mark)

Compute the regression line of height of the brother on the height of the sister. (Hint: do not use the original observations, but use the relationship between the correlation coefficient and the slope of a regression line i.e. $r = \frac{S_x}{S_y} b$ or equivalent: $b = \frac{S_y}{S_x} r$) (1mark)

   iii. What is the best prediction of the height of a brother of a sister with height 70 inches? (2 marks)
   iv. What is the estimated amount of variability in height of brother explained by height of sister? (1mark)
c. Test with a parametric method the hypothesis that the mean height of brother and sister is equal (2marks)
   i. Give the upper bound for the p-value (1mark)
   ii. Give the 95% confidence interval for the mean difference (1mark)

3. Question three (20 marks).

Human beta-endorphin (HBE) is a hormone secreted by the pituitary gland under the condition of stress. An exercise physiologist measured the resting (unstressed) blood concentration of HBE in three groups of men: 15 who had just entered a physical fitness program, 11 who had been jogging regularly for some time, and 10 sedentary people. The mean and standard deviations of the HBE levels (pg/ml) are shown in the table below. Calculations based on the raw data yielded SS (between)=240.69 and SS (within)=6,887.6.

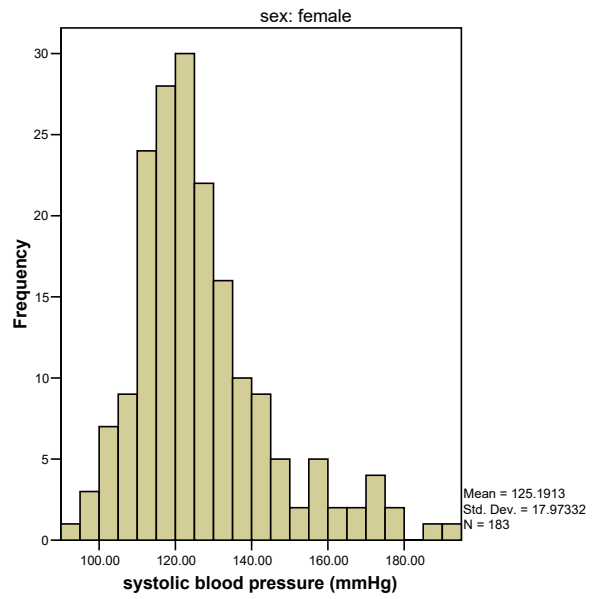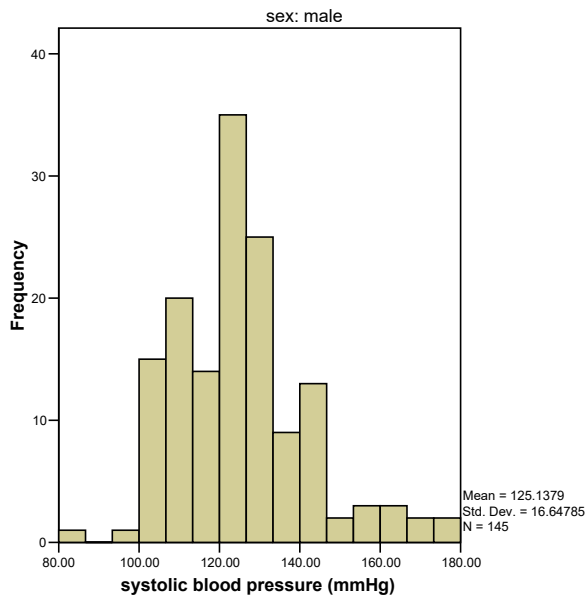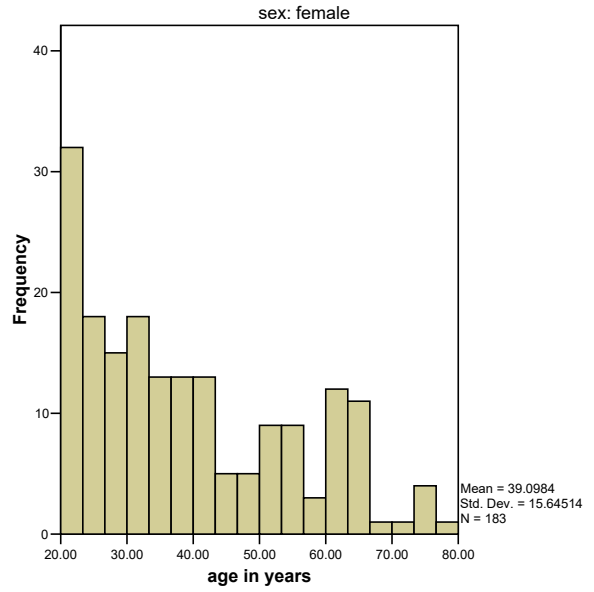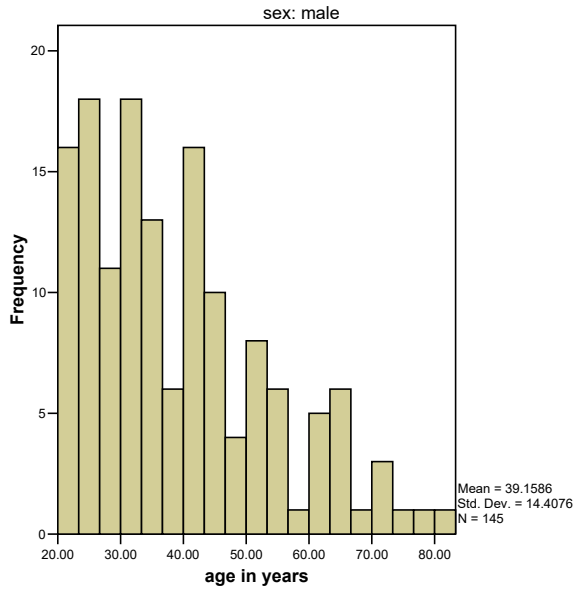|	Fitness program entrants	Joggers	Sedentary
Mean	38.7	35.7	42.5
SD	16.1	13.4	12.8
N	15	11	10

a. Make and complete the ANOVA table (10 marks)

b. Test the null hypothesis that there is no difference in mean HBE levels between the three groups (5 marks)
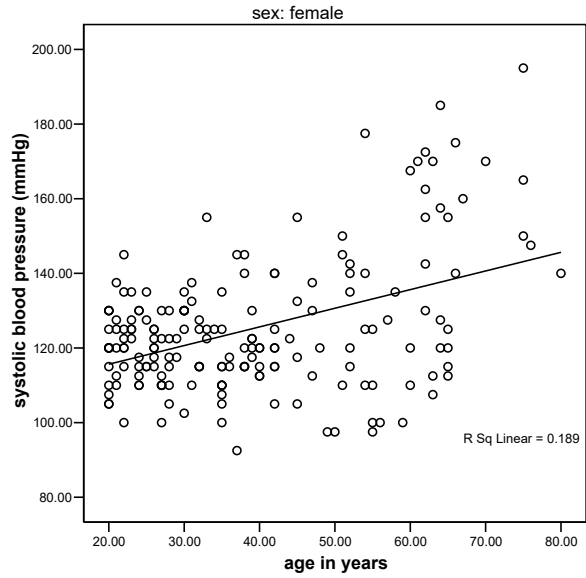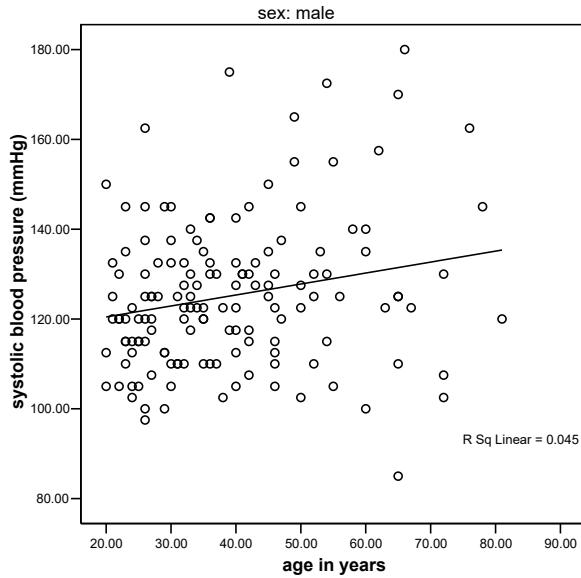c. What is the pooled standard deviation (5 marks)

4. **Question four (20 marks).**

A medical investigator selected from the population of some rural villages in a certain developing country 328 people for his study. Among other variables, systolic and diastolic blood pressure, body weight and pulse frequency were measured. Age and sex were also registered. In the accompanying SPSS output you will find some descriptive statistics and the results of the simple regression analyses of systolic blood pressure on age for males (sex=1, n=145) and females (sex=2, n=183) separately. Use this SPSS output to answer the following questions. First study the results of the analysis for the <u>females</u>. Questions (a) to (f) refer to this analysis.

Descriptive Statistics

sex		N	Minimum	Maximum	Mean	Std. Deviation
male	age in years	145	20.00	81.00	39.1586	14.40760
	systolic blood pressure (mmHg)	145	85.00	180.00	125.1379	16.64785
	Valid N (listwise)	145				
female	age in years	183	20.00	80.00	39.0984	15.64514
	systolic blood pressure (mmHg)	183	92.50	195.00	125.1913	17.97332
	Valid N (listwise)	183				

sex: male

Frequency (age in years)

Mean = 39.1586
Std. Dev. = 14.4076
N = 145

sex: female

Frequency (age in years)

Mean = 39.0984
Std. Dev. = 15.64514
N = 183

sex: male

Frequency (systolic blood pressure (mmHg))

Mean = 125.1379
Std. Dev. = 16.64785
N = 145

sex: female

Frequency (systolic blood pressure (mmHg))

Mean = 125.1913
Std. Dev. = 17.97332
N = 183

Model Summary

sex	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
male	1	.211(a)	.045	.038	16.32811
female	1	.435(a)	.189	.185	16.22763

a Predictors: (Constant), age in years

Coefficients(a)

sex	Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
			B	Std. Error	Beta		
male	1	(Constant)	115.569	3.939		29.340	.000
		age in years	.244	.094	.211	2.587	.011
female	1	(Constant)	105.649	3.237		32.642	.000
		age in years	.500	.077	.435	6.501	.000

a Dependent Variable: systolic blood pressure (mmHg)

a. Give the estimate for the mean systolic blood pressure of sixty year old women (2marks)

b. Give an estimate of the mean increase per age **decade** for the systolic blood pressure. (1mark)

   i. Give 95% confidence interval for it. (2marks)

c. What percentage of variability for the mean systolic blood pressure of 60 year old women? (1mark)

d. Give the 99% confidence interval for the mean systolic blood pressure of 60 year old women. (1mark)

e. One of the assumptions underlying the presented simple regression analysis is that the relation between age and systolic blood pressure is linear. Give two possibilities to check this assumption. (2marks)

f. From the histogram of the systolic blood pressure one can conclude that the distribution is not normal (the distribution is somewhat skewed to the right). Does this imply that the normality assumption underlying linear regression analysis is not fulfilled in this case? (1mark)

Now study also the results of the regression analysis for the <u>male</u>, and answer the following questions.

g. It will strike you that the correlation coefficient between age and systolic blood pressure is lower for males than for females. Is the difference statistically significant? (2marks)

h. Test whether the difference in mean **yearly** increase of the systolic blood pressure is significantly different between men and women. (the numbers are large, so use a simple and straightforward test). (3marks)

i. The difference in systolic blood pressure between men and women could be studied with the following multiple regression model.

$$systolic\ blood\ pressure = \beta_0 + \beta_1 age + \beta_2 Sex + \beta_3 Sex * age + residual$$

Using the accompanying regression analyses for men and women, give estimates of the β's in this model and their interpretations. Will age play the role of a confounder or effect modifier? (5marks)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	115.569	3.925		29.441	.000	107.846	123.292
	age in years	.244	.094	.212	2.596	.010	.059	.430
	sex	-9.920	5.093	-.284	-1.948	.052	-19.940	.100
	Age*Sex	.255	.122	.334	2.100	.037	.016	.495

a. Dependent Variable: systolic blood pressure (mmHg)

5. Question five (20 marks).

The data in the table were collected in Bradford, England, between 1968 and 1977, and relate to 13,384 women giving birth to their first child. The women were classified according to social class (five categories on the Registrar General's scale I-V) and according to the number of cigarettes smoked per day during pregnancy (on a three level categorization: 1 means no smoking, 2 means 1-19 cigarettes per day, and 3 means 20 or more cigarettes per day). The data for each category consist of counts of women showing toxaemic signs (hypertension and/or proteinuria) during pregnancy. The question of interest is how the toxaemic signs vary with smoking status, adjusted for social class. Some SPSS output is given below.

Social Class	Smoking category	No. of women w
1	1	
1	2	
1	3	
2	1	
2	2	
2	3	
3	1	
3	2	
3	3	
4	1	
4	2	
4	3	
5	1	
5	2	
5	3	

		Smoking category						Group Total	
		non smoking		1-19 cigarettes		20 or more cigarettes			
		Count	Col %	Count	Col %	Count	Col %	Count	Col %
Toxaemic category	No	5132	67.5%	3625	74.4%	658	72.8%	9415	70.3%
	Yes	2473	32.5%	1250	25.6%	246	27.2%	3969	29.7%
Group Total		7605	100.0%	4875	100.0%	904	100.0%	13384	100.0%

		Social class									Group Total	
		1		2		3		4		5	Count	Col %

		Count	Col %	Count	Col %	Count	Col %	Count	Col %	Count	Col %		
Smoking category	non smoking	417	77.4%	1135	71.2%	4703	56.8%	984	47.7%	366	40.4%	7605	56.8%
	1-19 cigarettes	105	19.5%	406	25.5%	3054	36.9%	859	41.6%	451	49.7%	4875	36.4%
	20 or more cigarettes	17	3.2%	52	3.3%	523	6.3%	222	10.8%	90	9.9%	904	6.8%
Group Total		539	100.0%	1593	100.0%	8280	100.0%	2065	100.0%	907	100.0%	13384	100.0%

				Toxaemic category		Group Total
				no	Yes	
				Count	Count	Count
Social class	1	smoking	non-smoking	286	131	417
			smoking	84	38	122
		Group Total		370	169	539
	2	smoking	non-smoking	785	350	1135
			smoking	318	140	458
		Group Total		1103	490	1593
	3	smoking	non-smoking	3160	1543	4703
			smoking	2683	894	3577
		Group Total		5843	2437	8280
	4	smoking	non-smoking	656	328	984
			smoking	812	269	1081
		Group Total		1468	597	2065
	5	smoking	non-smoking	245	121	366
			smoking	386	155	541
		Group Total		631	276	907

Risk Estimate

				95% Confidence Interval	
Social class			Value	Lower	Upper

		Value	Lower	Upper
1	Odds Ratio for Smoking (non-smoking / smoking)	.988	.639	1.526
	For cohort Toxaemic category = no	.996	.870	1.141
	For cohort Toxaemic category = Yes	1.009	.748	1.361
	N of Valid Cases	539		
2	Odds Ratio for Smoking (non-smoking / smoking)	.987	.780	1.249
	For cohort Toxaemic category = no	.996	.927	1.071
	For cohort Toxaemic category = Yes	1.009	.857	1.188
	N of Valid Cases	1593		
3	Odds Ratio for Smoking (non-smoking / smoking)	.682	.619	.752
	For cohort Toxaemic category = no	.896	.871	.921
	For cohort Toxaemic category = Yes	1.313	1.224	1.408
	N of Valid Cases	8280		
4	Odds Ratio for Smoking (non-smoking / smoking)	.663	.547	.802
	For cohort Toxaemic category = no	.888	.839	.939
	For cohort Toxaemic category = Yes	1.340	1.169	1.535
	N of Valid Cases	2065		
5	Odds Ratio for Smoking (non-smoking / smoking)	.813	.610	1.083
	For cohort Toxaemic category = no	.938	.858	1.026
	For cohort Toxaemic category = Yes	1.154	.947	1.406
	N of Valid Cases	907		

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymp. Sig. (2-sided)
Breslow-Day	11.551	4	.021
Tarone's	11.550	4	.021

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			.724
ln(Estimate)			-.324
Std. Error of ln(Estimate)			.039
Asymp. Sig. (2-sided)			???
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	???
		Upper Bound	???
	ln(Common Odds Ratio)	Lower Bound	???
		Upper Bound	???

a. First look at the cross table of toxaemic signs against smoking category. What test would you choose for the null hypothesis that there is no association between smoking and toxaemic signs? (1mark)
   i. What are the estimated risks of toxaemic signs for the three smoking categories? (1mark)
   ii. Give for the first smoking category also the corresponding standard error and 95% confidence interval. (1mark)
b. Compute the odds ratio of toxaemic signs of the combined second and third smoking category relative to the non-smokers. (1mark)
   i. Give also a 95% confidence interval (1mark)
c. Study the cross table of smoking category against social class. Do you think that smoking and social class are associated? (1mark)
   i. What would be an appropriate test(s) for the hypothesis that there is no association? (1mark)
d. To correct for possible confounding by social class, a stratified analysis was carried out using SPSS. Since the stratified analysis in SPSS cannot handle larger than 2X2-tables, the smoking categories were combined. Study the output and answer the following questions.
   i. Give the odds ratio per class and compare them with the overall unadjusted one. (1mark)
   ii. What is the most appropriate test for testing the hypothesis that there is no association between toxaemic signs and smoking, adjusted for social class? (1mark)
   iii. Give the value of the test statistic and p-value. (1mark)
   iv. Does possible heterogeneity of odds ratio across strata invalidate this test? (1mark)
   v. Is there statistical evidence for the odds ratios being not homogeneous across strata? (1mark)
   vi. Give the results of the tests and comment on it. (1mark)
   vii. What is the Mantel-Haenzel estimated odds ratio? (1mark)
   viii. What is its interpretation in this case? (1mark)
   ix. Fill in the question marks in the last table. (5 marks)