JARAMOGI OGINGA ODINGA UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF BIOLOGICAL, PHYSICAL, MATHEMATICS AND ACTUARIAL SCIENCE
UNIVERSITY EXAMINATION FOR DEGREE OF BACHELOR OF SCIENCE IN ACTUARIAL SCIENCE

3rd YEAR 2nd SEMESTER 2023/2024 ACADEMIC YEAR

MAIN REGULAR

COURSE CODE: WAB 2342

COURSE TITLE: TEST OF HYPOTHESIS

EXAM VENUE: **STREAM: (BSc. In Actuarial Science)**

DATE: **EXAM SESSION:**

TIME: 2.00 HOURS

Instructions:

i. Answer question one and any other two 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 a statistical software.
v. Unless otherwise stated conduct statistical tests at 95% confidence level

Assessment dataset

The High School and Beyond Data Set (HSB) is a dataset containing demographic information and standardized test scores of high school students (also available here (https://nces.ed.gov/surveys/hsb/). The dataset consists of 200 rows and 10 variables. The variables include the student's id, gender, ethnic background, socio-economic status, school type, program type, and scores from tests of reading, writing, math, science, and social studies. The dataset was collected as part of a large-scale longitudinal study conducted by the National Opinion Research Center in 1980 under contract with the National Center for Education Statistics. The students in the study are a nationally representative sample of n = 600 high school seniors with observations on 15 variables. The dataset will be provided.

QUESTION ONE (30 MARKS)

a) State three forms of data analysis (3 Marks)

b) The Delmonte company sells a large sterilizer with four extendable shelves for medical tools. Company engineers believe that the time to reach operating temperature from cold start (y, measured in minutes) is linearly related to the thickness of installation (x, in inches) A random sample of size $n = 12$ thickness was selected and time to reach operating temperature recorded for each. The data and summary statistics are as follows:

X	1.3	1.8	0.9	1.6	2.6	1.5	2.1	3.0	0.8	2.4	2.5	2.6
Y	8.0	6.9	8.1	7.0	6.3	6.5	6.4	5.8	8.3	8.3	6.6	6.6

$\sum x_1 = 23.1, \quad \sum y_1 = 84.8, \quad \sum x_1 y_1 = 1585\,, \quad \sum x_1^2 = 50.13\,, \quad \sum y_1^2 = 607.66$

i Obtain the regression parameter estimates and their standard errors (5 marks)

ii Write down the regression equation and interpret the results (4 marks)

iii Obtain a 95% confidence interval estimate for β_1 (3 marks)

iv Test the hypothesis H_0: β_1 = 0 vs H_1: $\beta_1 \neq 0$ at 5% level of significance (4 marks)

v Obtain the ANOVA table and compute F test for a significance regression. Use a significance level of 0.05 (4 marks)

a) Use hsb dataset to answer the following questions.

a. Test equality of two variances involving math and science showing hypothesis tested, test statistic, and significance. (3 marks)

b. Use appropriate test to evaluate equality of the two sample means and interpret your results (3 marks)

c. Sketch a scatter diagram to illustrate the relationship and interpret the output (4 marks)

QUESTION TWO (20 MARKS)

Use hsb dataset to answer the following questions.

i. Provide mean and corresponding standard deviations of math, science, read, write, and socst. (2 marks)

ii. Use an appropriate test to assess whether the average performance in math=55, its confidence interval and interpret (2 marks)

iii. Generate a variable that sums all the performance scores (read, write, math, science and socst). Name the new variable as "totalscore".

a) Develop a histogram with a density curve overlaid. Sketch and interpret the output. (2 marks)

b) Assess the normality of the totalscore with a stem and leaf plot and sketch the plot (2 marks)

c) Evaluate the distribution of totalscore across various races with a boxplot, sketch the plot and interpret (2 marks)

d) Use an appropriate graph to assess the direction of the relationship between total score(y) and math(x). (2 marks)

e) Provide intercept and slope parameters and interpret accordingly (2 marks)

f) Give confidence interval for the two regression parameters (2 marks)

g) Provide coefficient of determination and interpret (1 mark)

h) Develop an ANOVA table for the model and use it to test the hypothesis that $\boldsymbol{\beta_1}$=0 (1 marks)

i) Use the confidence interval obtained to test the hypothesis on the slope parameter $\boldsymbol{\beta_1}$. Compare and contrast (h) and (i) (2 marks)

QUESTION THREE (20 MARKS)

Use hsb dataset and generate a variable that sums all the performance scores (read, write, math, science and socst). Name the new variable as "totalscore" and where necessary use it to answer the following questions.

a. Differentiate between linear regression model (lm), generalized linear model (GLM), generalized additive models (GAM). (2 marks)

b. Fit a simple GLM with totalscore as dependence variable verses science as independent variable and provide the regression equation. Let this be the reduced model (**model1**). (3 marks)

c. Fit a multiple GLM with totalscore as dependence variable verses science, female, and schtyp as independent variables and write the regression equation and interpret the parameter estimates. Let this be the full model (**model2**). (3 marks)

d. Explain why R-squared may not be a good measure of fit statistic to this data. Suggest and deploy to the data an alternative fit statistic test for the model (3 marks)

e. Which model fit the data more accurately between reduced and the full model? Explain your answer. (3 marks)

f. Suppose it is suspected that violation of normality assumption may have contributed to inaccurate conclusion in (e). Deploy an appropriate corrective mechanism to the more accurate model in (e) and provide the new regression equation and fit statistic (3 marks).

g. A gain run a regression model with totalscore and independent variables science, female, schtyp and an interaction term (female*schtyp). Write the regression equation. Call this factorial model (**model3**). (2 marks)

h. Compare **model2** and **model3** with partial F test for ANOVA. (1 marks)

QUESTION FOUR (20 MARKS)

Use hsb dataset and generate a variable that sums all the performance scores (read, write, math, science and socst). Name the new variable as "totalscore" and where necessary use it to answer the following questions.

a. Develop a binary outcome on totalscore and call it "college" such that college=1, if totalscore>=280, otherwise college=0. Label the levels as either "Yes" or "No" and provide a table with counts and their corresponding percentage summaries. (2 marks)

b. Cross tabulate college and race and provide a table with count frequencies (2 marks)

c. Test the hypothesis whether college and race are independent, state the hypothesis and interpret the results (2 marks)

d. When is a Chi-square and Fisher's exact test used? (2 marks)

e. Develop a binary logit model involving college(y) and regressor variables (x variables) math, female, schtyp, and race. Write down the logit model. (2 marks)

f. Explain what the coefficients in a logistic regression tell us (i) for a continuous predictor variable and

(ii) for an indicator variable. (2 marks)

g. Interpret the regression coefficients taking into account reference categories (2 marks)

h. Obtain the odds ratio for the regression coefficients and interpret (2 marks)

i. Obtain fit statistic and interpret accordingly. (2 marks)

j. Predict the probability of high school and beyond pass to college for an Africa-American female student in public school with 60 marks in math. (2 marks)

QUESTION FIVE (20 MARKS)

Merge “daysabsent” data with the hsb dataset using the available common id and use the resulting dataset to answer the following questions.

a. Provide mean and standard deviation of “daysabsent” (2 marks)

b. Provide mean and corresponding standard deviations of “daysabsent” by race (4 marks)

c. State three assumptions necessary for a Poisson regression model to be considered (3 marks)

d. Assuming a Poisson process, fit a Poisson regression model involving daysabsent(y) and centred values for totalscore, female, schtyp, and race. Provide a table with regression slope estimates, standard errors, z-scores and corresponding probability values. (4 marks)

e. Interpret the regression slope estimates using exponentiated values of the average log counts (4 marks)

f. Conduct goodness of fit test to see if the model fits the data (3 marks)