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

2nd Year 1st SEMESTER 2023/2024 ACADEMIC YEAR
MAIN REGULAR

COURSE CODE: WAB 2209

COURSE TITLE: STATISTICAL COMPUTING I

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

DATE: **EXAM SESSION: Sep-Dec 2023**

TIME: 2.00 HOURS

Instructions:

i. Answer question **ONE** and any other two.
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. Candidates are advised to carry their personal computers with **R** and **ISwR** package installed beforehand.

QUESTION ONE

a) How would you check whether two vectors are the same if they may contain missing (NA) values? (Use of the identical function is considered cheating!) (2 marks)

b) If x is a factor with n levels and y is a length n vector, what happens if you compute y[x]? (1 mark)

c) Write the logical expression to use to extract girls between 7 and 14 years of age in the juul data set. (2 marks)

d) What happens if you change the levels of a factor (with levels) and give the same value to two or more levels? (2 marks)

e) Calculate the probability for each of the following events:

   i. A normally distributed variable with mean 35 and standard deviation 6 is larger than 42. (1 mark)

   ii. Getting 10 out of 10 successes in a binomial distribution with probability 0.8. (1 mark)

   iii. $X < 0.9$ when X has the standard uniform distribution. (1 mark)

   iv. $X > 6.5$ in a $\chi 2$ distribution with 2 degrees of freedom. (1 mark)

f) A rule of thumb is that 5% of the normal distribution lies outside an interval approximately ±2*s* about the mean.

   i. To what extent is this true? (1 mark)

   ii. Where are the limits corresponding to 1%, 0.5%, and 0.1%? (1 mark)

   iii. What is the position of the quartiles measured in standard deviation units? (1 mark)

g) For a disease known to have a postoperative complication frequency of 20%, a surgeon suggests a new procedure. He tests it on 10 patients and there are no complications. What is the probability of operating on 10 patients successfully with the traditional method? (2 mark)

h) If you make a plot like plot(rnorm(10),type="o") with overplotted lines and points, the lines will be visible inside the plotting symbols. How can this be avoided? (1 mark)

i) How can you overlay two qqnorm plots in the same plotting area? What goes wrong if you try to generate the plot using type="l", and how do you avoid that? (2 marks)

j) Plot a histogram for the react data set. Since these data are highly discretized, the histogram will be biased. Why? Consider truehist from the MASS package as a replacement. (2 marks)

k) Generate a sample vector z of five random numbers from the uniform distribution, and plot quantile(z,x) as a function of x (use curve, for instance). (2 marks)

l) In the data set vitcap, use a *t* test to compare the vital capacity for the two groups. Calculate a 99% confidence interval for the difference. The result of this comparison may be misleading. Why? (2 marks)

m) Perform the analyses of the react and vitcap data using nonparametric techniques. (2 marks)

n) Perform graphical checks of the assumptions for a paired *t* test in the intake data set. (1 marks)

o) The function shapiro.test computes a test of normality based on the degree of linearity of the Q–Q plot. Apply it to the react data. Does it help to remove the outliers? (2 marks)

QUESTION TWO [20 marks]

a) With the rmr data set, plot metabolic rate versus body weight. Fit a linear regression model to the relation. According to the fitted model, what is the predicted metabolic rate for a body weight of 70 kg? Give a 95% confidence interval for the slope of the line. (2 marks)

b) In the juul data set, fit a linear regression model for the square root of the IGF-I concentration versus age to the group of subjects over 25 years old. (2 marks)

c) In the malaria data set, analyze the log-transformed antibody level versus age. Make a plot of the relation. Do you notice anything peculiar? (2 marks)

d) One can generate simulated data from the two-dimensional normal distribution with a correlation of ρ by the following technique:

   i. Generate X as a normal variate with mean 0 and standard deviation 1. Sketch a histogram; (1 mark)

   ii. Generate Y with mean ρX and standard deviation sqrt(1 - ρ2). Use this to create scatterplots of simulated data with a given correlation. (2 marks)

   iii. Compute the Spearman and Kendall statistics for some of these data sets. (1 mark)

e) Do the values of the react data set (notice that this is a single vector, not a data frame) look reasonably normally distributed? Does the mean differ significantly from zero according to a *t* test? (2 marks)

f) In the data set vitcap, use a *t* test to compare the vital capacity for the two groups. Calculate a 99% confidence interval for the difference. The result of this comparison may be misleading. Why? (2 marks)

g) Perform the analyses of the react and vitcap data using nonparametric techniques. (2 marks)

h) Perform graphical checks of the assumptions for a paired *t* test in the intake data set. (2 marks)

i) The function shapiro.test computes a test of normality based on the degree of linearity of the Q–Q plot. Apply it to the react data. Does it help to remove the outliers? (2 marks)

QUESTION THREE [20 marks]

a) The zelazo data are in the form of a list of vectors, one for each of the four groups. Convert the data to a form suitable for the use of lm, and calculate the relevant test. Consider *t* tests comparing selected subgroups or obtained by combining groups. (3 marks)

b) In the lung data, do the three measurement methods give systematically different results? If so, which ones appear to be different? (3 marks)

c) Repeat the previous problem in (3b) using the zelazo and lung data with the relevant nonparametric tests. (2 marks)

d) The igf1 variable in the juul data set is arguably skewed and has different variances across Tanner groups. Try to compensate for this using logarithmic and square-root transformations, and use the Welch test. However, the analysis is still problematic — why? (3 marks)

e) A rule of thumb is that 5% of the normal distribution lies outside an interval approximately ±2*s* about the mean. To what extent is this true? Where are the limits corresponding to 1%, 0.5%, and 0.1%? What is the position of the quartiles measured in standard deviation units? (3 marks)

f) For a disease known to have a postoperative complication frequency of 20%, a surgeon suggests a new procedure. He tests it on 10 patients and there are no complications. What is the probability of operating on 10 patients successfully with the traditional method? (3 marks)

g) Simulated coin-tossing can be done using rbinom instead of sample. How exactly would you do that? (3 marks)

QUESTION FOUR [20 marks]

a) Create a factor in which the blood.glucose variable in the thuesen data is divided into the intervals (4, 7], (7, 9], (9, 12], and (12, 20]. Change the level names to “low”, “intermediate”, “high”, and “very high”. (3 marks)

b) In the bcmort data set, the four-level factor cohort can be considered the product of two two-level factors, say period and area. How can you generate them? (3 marks)

c) Convert the ashina data to the long format. Consider how to encode whether the vas measurement is from the first or the second measurement session. (2 marks)

d) Split the stroke data according to obsmonths into time intervals 0–0.5, 0.5–2, 2–12, and 12+ months after stroke. (2 marks)

e) The secher data are best analyzed after log-transforming birth weight as well as the abdominal and biparietal diameters. Fit a prediction equation for birth weight. How much is gained by using both diameters in a prediction equation? The sum of the two regression coefficients is almost exactly 3 — can this be given a nice interpretation? (4 marks)

f) The tlc data set contains a variable also called tlc. This is not in general a good idea; explain why. Describe tlc using the other variables in the data set and discuss the validity of the model. (2 marks)

g) The analyses of cystfibr involve sex, which is a binary variable. How would you interpret the results for this variable? (2 marks)

h) Consider the juul2 data set and select the group of those over 25 years old. Perform a regression analysis of √igf1 on age, and extend (2 marks)

QUESTION FIVE [20 marks]

a) Set up an additive model for the ashina data (see Exercise 5.6) containing additive effects of subjects, period, and treatment. Compare the results with those obtained from *t* tests. (2 marks)

b) Perform a two-way analysis of variance on the tb.dilute data. Modify the model to have a dose effect that is linear in log dose. Compute a confidence interval for the slope. An alternative approach could be to calculate a slope for each animal and perform a test based on them. Compute a confidence interval for the mean slope, and compare it with the preceding result.

(4 marks)

c) Consider the following definitions:

```
a <- gl(2, 2, 8)
b <- gl(2, 4, 8)
x <- 1:8
```

```
y <- c(1:4,8:5)
z <- rnorm(8)
```

Generate the model matrices for models z ~ a*b, z ~ a:b, etc. Discuss the implications. Carry out the model fits, and notice which models contain singularities. (4 marks)

d) Analyze the vitcap2 data set using analysis of covariance and draw your conclusions. Try using the drop1 function with test="F" instead of summary in this model, compare and contrast the results. (3 marks)

e) In the juul data set make regression analyses for prepubescent children (Tanner stage 1) of √igf1 versus age separately for boys and girls. Compare the two regression lines. (2 marks)

f) Try step on the kfm data and discuss the result. One observation appears to be influential on the diagnostic plot for this model — explain why. What happens if you reduce the model further? (2 marks)

g) For the juul data, fit a model for igf1 with interactions between age, sex, and Tanner stage for those under 25 years old. Explain the interpretation of this model. Hint: A plot of the fitted values against age should be helpful. Use diagnostic plots to evaluate possible transformations of the dependent variable: untransformed, log, or square root. (3 marks)