Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCE (SIMS)
MASTERS OF SCIENCE IN BIOMATHEMATICS
END OF SEMESTER EXAMINATION
BMA 8101: INTRODUCTION TO MATHEMATICAL COMPUTING

DATE: 19th April 2022 TIME: 3 Hours

INSTRUCTIONS
1. This examination consists of FOUR questions.
2. Answer Question ONE (COMPULSORY) and any other TWO questions.

QUESTION ONE (30 MARKS)

a) In a certain clinic the following measurements of 5 babies were taken;
Weight: 3.60, 2.93, 3.75, 4.17, 3.80

If the names associated with the above data are respectively: Smith, Claire, Joseph, Mary, Agatha then concatenate the weight and names to form vectors, and bind the columns to visualize the data. (4 marks)

b) Describe a data frame, and use the `as.data.frame()` function to construct one from a 4x3 matrix whose elements are 1,2,3,…12. (4 marks)

c) The data on admission to a certain university by gender in various departments is as given below;

| GENDER | DEPARTMENTS | | | | | |
|---|---|---|---|---|---|---|
| | Civil Eng. | Mech. Eng. | Hospitality | Business Studies | Nursing | Statistics |
| Male | 512 | 353 | 120 | 138 | 53 | 22 |
| Female | 89 | 17 | 202 | 131 | 94 | 24 |

Write a R code that utilizes `rbind()` function to do a chi-squared hypothesis test of homogeneity to investigate if the distribution of males admitted is similar to that of females. If the p-value is 2.2e-16 then what do you conclude? (5 marks)

d) A study tested whether cholesterol was reduced after using a certain brand of margarine as part of a low fat, low cholesterol diet. The subjects consumed on average 2.31g of the active ingredient, stanol easter, a day. The table below shows the cholesterol level (in mmol/L) of the study participants before and after 8 weeks;

| Subject ID | Before | After8weeks |
|---|---|---|
| 1 | 6.42 | 5.75 |
| 2 | 6.76 | 6.13 |
| 3 | 6.56 | 5.71 |
| 4 | 4.8 | 4.15 |
| 5 | 8.43 | 7.67 |

| | | |
|---|---|---|
| 6 | 7.49 | 7.05 |
| 7 | 8.05 | 7.1 |
| 8 | 5.05 | 4.67 |
| 9 | 5.77 | 5.33 |
| 10 | 3.91 | 3.66 |
| 11 | 6.77 | 5.96 |
| 12 | 6.44 | 5.64 |
| 13 | 6.17 | 5.51 |
| 14 | 7.67 | 6.96 |
| 15 | 7.34 | 6.82 |
| 16 | 6.85 | 6.29 |
| 17 | 5.13 | 4.45 |
| 18 | 5.73 | 5.17 |

i) Write a R code that depicts a histogram showing the amount of cholesterol levels after 8 weeks on diet in the study. (2 marks)

ii) Use the `boxplot()` function to write a R code that displays a horizontal boxplot of the distribution of cholesterol after 8 weeks. (2 marks)

iii) If healthy human cholesterol after 8 weeks is described using a Normal model with the mean and standard deviation found in the sample, then use the `pnorm()` function to write a R script to determine what proportion of healthy individuals would be expected to have a cholesterol level less than 5.67 mmol/L? (3 marks)

e) State how to best summarize (for center and spread) data that is not Normally distributed. (2 marks)

f) Describe an outlier using the interquartile range. (3 marks)

g) Extend the following R script using the `subset()` and `anti_join()` functions to create and view a dataset called `iris.modified` that is devoid of the `virginica` species;

```
library("dplyr")
data("iris")
```
(5 marks)

## QUESTION TWO (20 MARKS)

Suppose you have a dataset named `Birthweight` that contains information on new born babies (weight and length) and their mothers (age and gestation period). Write a R code that does the following tasks;

a) Imports the `Birthweight` dataset from Documents folder in your PC into RStudio, and assign it the name `birthweight`. (2 marks)

b) Create a well-labeled scatterplot to display the relationship between gestation period and baby weight. (3 marks)

c) Calculate the correlation between gestation period and weight. (2 marks)

d) To determine the regression equation that relates gestation period and baby weight. (2 marks)

e) Modify the scatterplot from a) above to include the regression line (red in color) and correlation information. (3 marks)

f) Create a residual plot for the gestation period and label it accordingly. (3 marks)

g) Use the regression line to determine the expected weight for a baby whose mother gestation period 40 weeks. (3 marks)

h) Displays $R^2$ for this relationship. (2 marks)

**QUESTION THREE (20 MARKS)**

a) Consider the following contingency table of heart attack status by the type of treatment (placebo versus aspirin);

| | Heart attack | No heart attack | Total |
|---|---|---|---|
| **Placebo** | 189 | 10845 | 11034 |
| **Aspirin** | 104 | 10933 | 11037 |
| **Total** | 293 | 21778 | 22071 |

i. Determine the sample proportion of people, P1 and P2, who suffered from a heart attack in the placebo and aspirin groups respectively. Which people have a lower risk of heart attack? (3 marks)

ii. The sample proportions are, in this case, related to the risk of heart attack. Therefore, determine the relative risk of suffering from heart attack and interpret your results. (2 marks)

iii. Compute the odds of a heart attack in the placebo group, O1, and in the aspirin group, O2. Then obtain the odds ratio and interpret the results. (4 marks)

b) A dataset called `Mosquito.xlsx` contains data recorded in an experiment conducted on male soldiers in the Indian Army who were stationed in the Tezpur/Solmara garrison in Northeast India. Thirty soldiers were randomly selected to receive one of five types of mosquito repellant patch. Three of the treatments were a single repellant and two were combinations of two repellants. After giving informed consent, the study participants affixed the patches at predetermined points on their uniforms and research assistants (who were blinded to the type of repellant used) counted the number of times a mosquito landed on each individual in an hour. The aim was to determine if there is a difference in the mean number of mosquito landings between soldiers who wore patches with a single repellant and soldiers who wore patches with a combination of two repellants.

i. State why this is an experimental design in which we are comparing two independent means. (2 marks)

ii. Identify the null and alternative hypotheses. (4 marks)

iii. State why t-test is the appropriate statistical test for these hypotheses. Hence, verify that the assumptions for using that test are met. (3 marks)

iv. Given the test results as below; What can you conclude about the mean number of mosquito touches? Interpret the 95% confidence interval for the difference in the mean number of mosquito touches between the two groups of soldiers. (2 marks)

```
                Welch Two Sample t-test
data: Mosquito$Mosq_count[Mosquito$Treatment == 0]
 and Mosquito$Mosq_count[Mosquito$Treatment == 1]
   t = 3.9539, df = 132.44, p-value = 0.0001246
alternative hypothesis: true difference in means is not
                      equal to 0
          95 percent confidence interval:
                 1.038327 3.117228
                 sample estimates:
       mean of x mean of y 8.011111 5.933333
```

**QUESTION FOUR (20 MARKS)**

The data below was collected from women of Pima Indian heritage living in Arizona, USA with the aim of testing for diabetes. The dataset is referred to as `Pima`. Each row corresponds to an individual in the sample, while each column gives the variable of interest. The description of the variables is as follows;

| | |
|---|---|
| npreg: | number of pregnancies. |
| glu: | plasma glucose concentration in an oral glucose tolerance test. |
| bp: | diastolic blood pressure. |
| skin: | triceps skin fold thickness (mm). |
| bmi: | body mass index. |
| ped: | diabetes pedigree function. |
| age: | age in years. |
| type: | disease status; Yes for diabetic and No for nondiabetic. |

| | npreg | glu | bp | skin | bmi | ped | age | type |
|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 86 | 68 | 28 | 30.2 | 0.364 | 24 | No |
| 2 | 7 | 195 | 70 | 33 | 25.1 | 0.163 | 55 | Yes |
| 3 | 5 | 77 | 82 | 41 | 35.8 | 0.156 | 35 | No |
| 4 | 0 | 165 | 76 | 43 | 47.9 | 0.259 | 26 | No |
| 5 | 0 | 107 | 60 | 25 | 26.4 | 0.133 | 23 | No |
| 6 | 5 | 97 | 76 | 27 | 35.6 | 0.378 | 52 | Yes |

Write R codes to perform the following tasks;

a) Create a new variable called `type.num` in the `Pima` dataset that is coded as 0 if the individual was nondiabetic and 1 if the individual was diabetic. (3 marks)

b) Calculate the frequency and relative frequency for the `type` variable. (2 marks)

c) The frequency histogram for the numerical variable `bmi`. (2 marks)

d) According to Centers for disease Contorl (CDC) the standard weight status based on BMI is;

| BMI | Weight Status |
|---|---|
| Below 18.5 | Underweight |
| 18.5 – 24.9 | Normal |
| 25.0 – 29.9 | Overweight |
| 30.0 and above | Obese |

Add a categorical variable `weight.status` based on the `bmi` variable in `Pima` dataset, using `if - else()` statements within a `for()` loop to produce the output below; (7 marks)

```
  npreg  glu   bp  skin bmi  ped  age type weight.status
1   5    86    68  28   30.2 0.364 24  No     Obese
2   7    195   70  33   25.1 0.163 55  Yes    Overweight
3   5    77    82  41   35.8 0.156 35  No     Obese
4   0    165   76  43   47.9 0.259 26  No     Obese
5   0    107   60  25   26.4 0.133 23  No     Overweight
6   5    97    76  27   35.6 0.378 52  Yes    Obese
```

e) The joint distribution of two categorical variables is displayed using a contingency table. Describe joint, marginal and conditional distributions with regard to a contingency table. (6 marks)