Auckland University of Technology

COMP616/STAT604 Assignment 2

Semester 1, 2024

Due Date: 2 June 2024 (11:59 pm) **Total Marks: 80**

Outline: This assignment has three questions and contributes 25% to your final grade.

Purpose: To assess your analytics and computing skills on the topics covered in Weeks 6-11.

Instructions:

1. **Use the Answer Booklet (.Rmd file) provided to type in your answers.** Enter your name and Student ID in the `author` field.
2. Save the file on disk. The file name must include 1) your last name, 2) your first name, and 3) your student id, e.g., if Jane Doe submits her assignment, her file must be named "Doe_Jane_123456789''.
3. Finally, knit your .Rmd answer-file to .pdf and submit the PDF to Canvas.
4. Fill in and sign an assessment cover-sheet which must be the very first page in the PDF. Use, e.g., Adobe Acrobat Pro on Uni computers. You can submit this cover sheet as a separate file.

IMPORTANT: Any `R` code employed to complete this assignment must be self-explanatory and must be embedded in your answer using R Markdown code chunks. Screenshots and other images will not be allowed and will be penalized. *Make sure you make comments on your code for a full understanding of your answer.*

Late submissions: There is a lateness penalty of 5 marks/day (or part of a day thereof), up to a maximum of 3 days, unless the student gets an approved SCA by NO later than 2 June. See note below.

Note: If you need an extension with no lateness penalty because, e.g., your performance has been impacted by some extenuating, unexpected, circumstances, you can submit and SCA along with relevant evidence using the submission link from the course home page. **Bear in mind that SCA processing may take up to 5 working days**. If you have questions, contact victor.miranda@aut.ac.nz or ryan.ip@aut.ac.nz.

Question 1 [25 marks]

(a) *(5 marks)* The following table summarises the height and weight of 10 randomly selected 4-year-old girls. Use R to find the *Spearman's rank correlation* and the *Kendall's tau-b*. Show your working by providing the R code used. *[Hint: Check lecture slides for Weeks 9 and 10.]*

Girl	1	2	3	4	5	6	7	8	9	10
Weight (kg)	19.0	17.2	16.7	17.0	17.4	12.4	10.1	14.6	13.4	14.4
Height (cm)	101.0	113.2	102.9	80.4	96.9	108.7	97.2	98.4	104.0	81.9

(b) *(5 marks)* Suppose you are conducting a *sign test* at the 5% level of significance to check if the median of a population is 55 or not. In a random sample of 12 observations, 2 were below 55, and 10 were above 55. Find the p-value manually and verify your answer using R. Hence, draw your conclusion. *[Hint: Check lecture slides for Weeks 6 and 8.]*

(c) *(5 marks)* A bag contains ten balls, four are blue and the rest are red. The balls are drawn one by one *without* replacement. Suppose you wish to determine if the sequence of the colours of the balls is random or not, answer the following questions:

- Which hypothesis test is suitable? Briefly explain your choice.
- What are the minimum and maximum values of the test statistic concerning the test selected? *[Hint: Check lecture slides for Weeks 6 and 8.]*

(d) *(5 marks)* Below shows the UV indices measured in two locations A and B. It is hypothesised that the UV index at location A is higher than that at location B. Use R to conduct the exact Wilcoxon Sum Rank Test at the 5% level of significance and draw your conclusion. *[Hint: Check lecture slides for Weeks 9 and 10.]*

```
locA <- c(6,5,4,3,7,8)
locB <- c(3,2,7,4,6,4)
```

(e) *(5 marks)* Suppose 10% of the population are contracted with COVID at any given time. In the clinical trials run by a particular brand of test kit, out of all people who had COVID, 90% were correctly identified as "positive". Meanwhile, out of all people who did not have COVID, 95% were correctly identified as "negative". Suppose you took a test and the test result was positive, what is the probability that you have COVID? *[Hint: Check lecture slides for Week 11.]*

Question 2 [29 marks]

Staff from four departments at AUT were randomly selected to sit a mathematics test. The test results (out of 60) are provided in the table below:

Department	Test Results
Art and Design	41, 52, 32, 44, 45, 49
Education	43, 60, 55, 53, 48
Midwifery	26, 27, 46, 51, 17
Law	30, 23, 39, 37, 13, 29

The aim of the study was to determine if there is a difference in the median mathematical skills among the staff from the above four departments. In answering the questions below, show your R code whenever possible.

(a) *(4 marks)* Choose a non-parametric method that is suitable for the purpose. Provide a brief explanation for this selection.

(b) *(4 marks)* Write down the null and alternative hypotheses.

(c) *(4 marks)* Enter the dataset into R in an appropriate format for further analysis. Hence, print the dataset.

- **Hint:** Check the labs and lecture slides for examples of the appropriate format required.

(d) *(4 marks)* Find the sum of ranks for the staff from the Midwifery department (with test results 26, 27, 46, 51, 17).

(e) *(5 marks)* Use R to conduct the test at the 5% level of significance and draw your conclusion.

(f) *(8 marks)* Conduct pairwise comparisons between the departments. Hence, with an aid of a suitable graph, determine the department(s) that has/have staff with the best mathematical skills.

Question 3 [26 marks]

In a study, the sizes of abalones caught in three areas were categorised as either "large", "medium" or "small". The findings are shown below:

Size	Area A	Area B	Area C	**Total**
Large	150	57	14	221
Medium	96	45	38	179
Small	52	25	23	100
—	—	—	—	—
Total	298	127	75	**500**

The aim of the study was to determine if there are any associations between the size of abalones and the areas. In answering the questions below, show your R code whenever possible.

(a) *(4 marks)* Choose a non-parametric method that is suitable for the purpose. Provide a brief justification for this choice.

(b) *(4 marks)* Write down the null and alternative hypotheses.

(c) *(4 marks)* Enter the table above into R in an appropriate format for further analysis. Hence, print the table.

- **Hint:** Check the labs and lecture slides for examples of the appropriate format required. Note that the row and column of totals do not need to be specified.

(d) *(5 marks)* Use R to find the column proportions. Hence, comment on the expected outcome in terms of the hypothesis test you selected in Parts (a) and (b) without computing the p-value. Just by looking at the proportions.

(e) *(4 marks)* Under the assumption that H_0 is true, manually compute the expected number of large and medium abalones in Area C.

(f) *(5 marks)* Use R to conduct the test at the 5% level of significance and draw your conclusion.

END OF ASSIGNMENT 2