401077 Introduction to Biostatistics, Autumn 2021

**Assignment 1**

**Due Sunday March 28, 2021**

Please answer each question in the template document provided and submit via Turnitin on or before the due date. The marks allocated to each question are shown in the assignment. A total of 30 marks are available and this assignment is worth 30% of your overall grade.

Many of the questions in this assignment ask you to conduct analyses on the assignment data set assigned to you. Read ‘Description of your data set.docx’ for the descriptions of the variables. The assigned data set is different for each student.

This assignment is assessing your skills, not the skills of the computer. You will need to include graphs in your assignment but all other R Commander output attracts 0 marks and is discouraged. It is your task to identify the relevant results in the R Commander output and write these up in your assignment.

Some of the assignment questions ask you to justify your answer or explain why. Answers without these requested justifications and explanations will be awarded 0 marks.

Question 1: (6 marks)

Consider the following questionnaire.

1. What is your gender? \_\_\_ Male \_\_\_Female
2. How old were you at your last birthday? \_\_\_Years
3. What grade do you expect to get in this Unit? \_\_\_ HD \_\_\_ D \_\_\_ C \_\_\_ P \_\_\_ F

Separately for each of the 3 questions in the questionnaire

1. identify the type of variable; and
2. identify the most appropriate graph to display the distribution of this variable

Question 2: (7 marks)

Consider the research question: What, if any, association exists between myopia and hours of sedentary activity per week in the sample of young adult Australians?

Answering this question requires you to analyse the assignment data set assigned to you. Only the variables myopia status ‘myopia’ and sedentary hours per week ‘sed’ should be used during this analysis.

To address this research question

1. Using R Commander graph the relationship between myopia status and sedentary hours per week. (1 marks)
2. Using R Commander calculate descriptive statistics for the centres of the distributions of sedentary hours per week for the myopia and normal samples separately. Name each statistic you are using and present its value. R Commander output alone is insufficient and will attract 0 marks. (1 mark)
3. Repeat b. for descriptive statistics of the spreads of the distributions. (1 mark)
4. Repeat b. for descriptive statistics of the shapes of the distributions. (1 mark)
5. Drawing upon the results in parts a, b, c and d above present an answer to the research question in one or two paragraphs. (3 marks)

Question 3 (9 marks)

This question requires you to analyse the assignment data set assigned to you. Only the variables myopia status ‘myopia’ and highest education level achieved ‘educ’ should be used during this analysis.

1. Use R Commander to tabulate the relationship between myopia and highest education level achieved in the sample. Type this table yourself with appropriate headings – R Commander output is insufficient and will attract 0 marks. Include both counts and row percentages in your table. (2 marks)
2. If you were to select one person at random from this data set, what is the probability they would have myopia? (1 mark)
3. If you selected one person with myopia at random from this data set what is the probability they have completed tertiary education? (1 mark)
4. Are myopia and education independent? Explain why or why not. (2 mark)
5. If you were to draw 10 people at random from this data set, what is the probability of that your sample will contain 4 or less people with myopia? To answer this question, you first need to describe how this probability can be obtained. After that you can use R Commander to actually calculate the answer. (3 marks)

Question 4 (3 marks)

Do you believe that the Normal model would provide a good model for the distribution of age in your assignment data set? Explain why or why not. Justify your answer with appropriate graphs and statistics. (3 marks)

Question 5 (5 marks)

The distribution of blood cholesterol level in the population of young adults aged 20 to 30 years is close to Normally distributed, with mean **= 172 milligrams per decilitre (mg/dl) and standard deviation **= 36 mg/dl.

1. Selecting one of the young adults at random, what is the probability that their blood cholesterol level is between 160 and 170 mg/dl? (1 mark)
2. You measure the blood cholesterol level of 100 young adults chosen at random and calculate the sample mean. What is the probability that this sample mean blood cholesterol level is between 160 and 170 mg/dl? Explain how you derive your answer. After that you can use R Commander to actually calculate the answer. (2 marks)
3. If you were to repeat your sample of 100 young adults 20 times, and record the sample mean for each, what would you expect the mean and standard deviation of the distribution of the sample means to be? Explain why. (2 marks)