DATA101 Assignment 3, Semester 2 2025

Due Date: 8th August 2025

Student ID Number: _____

Total Marks: 20

Instructions:

- Complete all questions in R Markdown format
- If using this .Rmd file for your solution, please enter your student ID number in the space provided above
- Submit your completed assignment as a PDF through LEARN by 11:59pm on the due date
- Show all your R code and explain your reasoning where required
- Code that produces errors will receive minimal marks

Question One: Sampling bias identification (4 marks)

A public health researcher wants to study the effectiveness of a new exercise program on reducing stress levels among office workers in Christchurch.

a. For each of the following sampling scenarios, identify whether sampling bias is present. If bias exists, explain what type of bias it is and how it might affect the results. (2 marks)

   Scenario A: The researcher posts flyers at local gyms asking for volunteers to participate in the study.

   Scenario B: The researcher randomly selects 5 office buildings in Christchurch, then surveys all workers in those buildings about their current stress levels before implementing the exercise program.

   Scenario C: The researcher obtains a list of all registered businesses in Christchurch and randomly selects 200 office workers from this list, but only 60% of those contacted agree to participate.

b. Suggest an improved sampling method that would minimise bias for this study. Justify your recommendation. (2 marks)

Question Two: Observational vs. experimental studies (6 marks)

A researcher wants to investigate the relationship between vitamin D levels and depression in adults.

a. Design a case-control study to investigate this relationship.

   Clearly describe:

   - How you would define cases and controls
   - What information you would collect from participants
   - How you would measure the association (2 marks)

b. Design a cohort study to investigate this relationship.

   Clearly describe:

   - How you would select your cohort
   - How long you would follow participants

- What outcomes you would measure (2 marks)

c. Compare the advantages and disadvantages of your case-control study versus your cohort study for investigating this research question. Which design would you recommend and why? (2 marks)

Question Three: Confounding and causality (5 marks)

A study found that people who drink coffee regularly have lower rates of liver disease compared to non-coffee drinkers.

a. Explain what confounding means in the context of epidemiological studies. Provide the three criteria that a variable must meet to be considered a confounder. (2 marks)

b. Identify two potential confounding variables that could explain the observed association between coffee consumption and liver disease. For each confounder, explain how it could be related to both coffee consumption and liver disease risk. (2 marks)

c. The researchers want to establish whether coffee consumption actually causes a reduction in liver disease risk. Explain why observational studies alone cannot definitively establish causality, and describe what type of study design would provide the strongest evidence for a causal relationship. (1 mark)

Question Four: Introduction to Probability (5 marks)

In Question Three, you used a study that found people who drink coffee regularly have lower rates of liver disease compared to non-coffee drinkers to discuss the potential for confounding in observational studies. In this question, we will present some data on the study to test your knowledge of probability. The data is contained in `coffeeliver.csv`, available on LEARN.

a. Import the data into R, and create a table using the `table` function summarising the information in the `Coffee` and `liver` variables. (1 mark)

Hint: As the data is stored as a .csv file, read the function `read.csv` *to import the data, with header=TRUE, as the file has column names* `Coffee` *and* `Liver`*.*

b. What are the set of possible outcomes (events) in the Coffee liver disease study? (1 mark)

Hint: You might find it easier to express the individual outcomes as combinations of events

c. From the information provided in the table you constructed in part a., what are your estimates of: (2 marks)

- Pr(Drinking Coffee)
- Pr(Having Liver Disease)
- Pr(Drinking Coffee or Having Liver Disease)

d. Are *Drinking Coffee* and *Having Liver Disease* examples of

- disjoint events,
- complete exhaustive events?

Explain why. (1 mark)