DATA101 Assignment 3, Semester 2 2025

Due Date: 8th August 2025

Student ID Number: 83972706

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. *Selection Bias - Any changes will be minial since the people at the gym are already doing their own exersise programs!*

   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.

Minimal bias, this is the ideal sampling procedure

`Scenario C: The researcher obtains a list of all registered businesses in Christchurch and randomly sel`

Again pretty good sample selection but there is a little bit of Non-Response Bias. The 40% who declined to participate might have different views, work habits, or characteristics than the 60% who did.

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

To make sure we're properly representing key groups of office workers like different age ranges, genders, industries, and company sizes I'd use stratified random sampling. That way, we're picking people randomly within each group, instead of just grabbing whoever's easiest to reach. It helps cut down on selection bias and makes sure we're getting a fair mix of Christchurch's office worker population. It also makes the results more reliable because they actually reflect the diversity of the people we're studying, not just whoever happened to volunteer.

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)

Is there an association between vitamin D levels and depression in adults

Cases = Adults diagnosed with depression

Controls = Adults without depression, matched by age and gender where possible.

I would collect the amount vitamin D that they have (Via blood test or something)

To measure the association I would use odds ratio compare the odds of low vitamin D levels in depressed individuals versus non-depressed individuals.

The odds ratio 'table' will look like: Horizontal line: Low vitamin D, Normal Vitamin D, Total and then the Vertical line: Depressed, Not depressed

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)

Cohort Selection: Recruit a group of adults with depression.

Measure their vitamin D levels (via blood tests) and categorize them (e.g. low, normal, high).

Participants will be followed for 1 year, with the depression serverity measured each month

Outcomes Measured: Depression severity

Measure severity using standardized depression scales.

Track changes in vitamin D levels.

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)

While each has their own pros and cons, the case-control study is the better option. Having both adults with depression, and adults without depression allows us to see the effects of vitamin D a lot better than just doing a single group like the cohort study. The cohort study would be better for large studies that are likley to continue for multiple years but overall I would suggest the case-control study.

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) *Confounding is when the outcome or results are influenced by a third variable*

1. *Must be associated with the independent variable*
2. *Must be associated with the dependent variable*
3. *It should not be a result of the independent variable (I can't really describe it simply, but an example would be if you were studying correlation between alcohol and lung disease or something, it's likely that people are drink are more likely to smoke than people who do not thus smoking is a confounding variable)*

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)

- *Smoking People who drink coffee are likely prone to the addictive effects of the caffeine. Addictive habbits such as that are often grouped with similar, such as smoking (Or even vaping) for the nicotine. Smoking will almost certainly have a strong effect on someone's liver.*
- *Alcohol Consumption People who do not drink coffee are far less likely to consume alcohol as well. One of the greatest risks of alcohol is the damage it does to your liver so this is obviously going to affect the results*

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) *People who drink coffee likely have different lifestyles to those who do not, for example they may not consume energy drinks which are usually high in sugar and other products that probably have an effect on the liver. A better study method would be Randomized Control Trial. Participants would be assigned to either a group that drinks coffee reguarly and a group that does not (And also avoids alternative caffeine sources such as energy drinks). This will help mitigate the effects of confounding variables.*

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)

```
csv_data <- read.csv('coffeeliver.csv', header = TRUE)

# I had to add all the labels and stuff the data was too hard to understand otherwise!!
# I think it will probably go off the page because of how long the line is.
summary_table <- table(factor(csv_data$Coffee, levels = c(0, 1), labels = c('Not Coffee', 'Coffee')), fa
summary_table
```

```
##
##              Not Liver Liver
##   Not Coffee      2341   374
##   Coffee          5184   461
```

```
*Hint: As the data is stored as a .csv file, read the function* \texttt{read.csv} to import the data, wi
```

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

Drinks coffee AND has liver disease

Drinks coffee AND does not have liver disease

Does not drink coffee AND has liver disease

Does not drink coffee AND does not have liver disease

```
*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)

```
# These are the values from the table in the last question
total <- 2341 + 374 + 5184 + 461

pr_coffee = (5184 + 461) / total
pr_liver = (374+461) / total
pr_coffee_liver = (5184 + 461 + 374) / total

pr_coffee
```

```
## [1] 0.6752392
```

```
pr_liver
```

```
## [1] 0.09988038
```

```
pr_coffee_liver
```

```
## [1] 0.7199761
```

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

- disjoint events,
- complete exhaustive events?

Explain why. (1 mark) *It is a complete exhaustive event. If they were disjoint events they would not be able to happen at the same time but we can clearly see in the table that they do sometimes occur together.*