# PHCM 9795: Foundations of Biostatistics

## Assignment 2

## Due: 11:59pm Sunday August 15, 2021

## Value: 40% of course assessment

## Instructions

Answer all questions.

## Formatting your assignment

- Submit your assignment online as one document in PDF format. This format locks the information and is particularly important to preserve the formatting, graphs and Stata output;

- Number every page and include your zID in the header of every page;

- Clearly number the answer to each question; only include your answer (not the question);

- Use 11 point or 12 point font and 1.5 times spacing (not double spacing);

- Format figures and tables appropriately according to the guidelines given in this course;

- Do not submit Stata data files.

## Submitting your assignment

Submit your assignment online through the course Moodle site by 11.59PM (AEST) Thursday August 12. Instructions about how to submit your assignment are provided in the Course Outline and on the Moodle site (see Assessments).

You must agree that the work is entirely your own before you can submit your assignment. You will be submitting your assignment via Turnitin, which is a similarity detection program.

## Late submission, collusion and plagiarism

Late submission without the permission of the course convenors will result in a penalty of two percent (2%) per day. Workload is usually not considered a valid reason for requesting an extension, nor is failure to save work and other electronic problems. Extensions for health-related reasons will be given with appropriate evidence.

You must submit your own work. All assignments will be electronically checked against all other assignments submitted for this course. Evidence of collusion constitutes Academic Misconduct and will be investigated. If proven, the grade of 0% will be awarded and your name will be placed on the University Academic Misconduct Register.

## Introduction

This assignment assesses your understanding of topics discussed in Modules 1 to 10 of the course.

It is important to present your answers to all questions with appropriate precision (i.e. in tables, figures and text). Tables and figures should be presented in the way you would present them in a scientific journal or a standard report. You should follow the presentation guidelines you have been given in the course. When asked to write a conclusion or interpret values, your answer should include relevant values you have calculated and other data provided to you.

## Question 1 [10 marks]

The six-minute walk test is designed to assess people's aerobic capacity and endurance. A person who undertakes the test is asked to walk as far as they possibly can on a hard, flat surface in six minutes. The distance the person walks is called the six-minute walk distance (6MWD). Previous studies have estimated that the 6MWD in people with chronic obstructive pulmonary disease (COPD) has an average of around 360 metres, with a standard deviation of 100 metres.

A researcher is interested in comparing the 6MWD in a group of people with a respiratory disease who undertake an exercise program compared with those who do not undertake the exercise program.

a) The researcher wishes to undertake a randomised controlled trial to study the effect of the intervention. The researcher believes the minimum clinically important difference is 35 metres. The desired power of the study is 90% and alpha = 0.05. What sample size would you recommend that she use? Include your Stata output in your answer. [3 marks]

b) A colleague suggests that a minimum clinically important difference should be set to 25 metres. Without performing any calculations, explain what will happen to the required sample size and why. Which scenario would you recommend: 25 or 35 metres and why? [3 marks]

c) Research in this area has suggested that around 20% of invited participants fail to attend their scheduled 6MWD. How many participants should the researchers enrol to meet the requirements of part (a)? [2 marks]

d) The researchers decide on scenario (a). After completing recruitment, the researchers found that 35% of the invited participants (calculated in part (a)) failed to attend. Calculate the revised power to detect a 35 metre difference in 6MWD. [2 marks]

## Question 2 [9 marks]

Researchers conducted a study to assess the effect of an education campaign in increasing knowledge of Treatment as Prevention as a method of preventing transmission of HIV. Each participant's knowledge was assessed before and after a brief education campaign. Knowledge was assessed using a seven-point scale where 1 represents low knowledge and 7 represents high knowledge.

The dataset *knowledge.dta* contains the following variables:

before: knowledge score before the education campaign
after: knowledge score after the education campaign

a) State the null hypothesis and alternative hypothesis for this study. [1 mark]

b) What test is appropriate to test your null hypothesis? Why is this test appropriate? Provide supporting figures and/or statistics to justify your decision. [2 marks]

c) Use Stata to conduct a test to test your null hypothesis. Copy and paste the Stata output into your assignment. [2 marks]

d) Write a brief conclusion summarising your results. [4 marks]

## Question 3 [9 marks]

A study was undertaken in a hospital to compare the performance of a new surveillance method to the standard method in the ability to detect patients who develop an infection after a caesarean section.

Individual record data from this study is provided in the file called *surveillence.dta*. The dataset contains the following variables:

ID: Patient identification number
new: Assessment of cases by the new surveillance method (1=Positive, 2=Negative)
standard: Assessment of cases by the standard method (1=Positive, 2=Negative)

a)  State the null hypothesis and alternative hypothesis for this study. [1 mark]

b)  Construct a 2 × 2 table to report the data provided. [2 marks]

c)  Use Stata to conduct a test to test your null hypothesis. Copy and paste the Stata output into your assignment. Why is this test appropriate? [2 marks]

d)  Write a brief conclusion summarising your results. [4 marks]


## Question 4 [12 marks]

A study was conducted to assess the association between oxygen intake rate (a measure of aerobic fitness) and a range of body size and exercise tests. Data from 186 men are stored in the dataset *fitness.dta*. The dataset contains the following variables:

Age (years)
Weight (kg)
Oxygen - oxygen intake rate (ml per kg body weight per minute)
RunTime - time to run 1.5 miles (minutes)
RestPulse - resting pulse (beats per minute)
RunPulse - heart rate while running (beats per minute)
MaxPulse - maximum heart rate recorded while running (beats per minute)

In this analysis, the researchers are interested in predicting oxygen intake rate (Oxygen: the outcome variable) from the time taken to run 1.5 miles (RunTime: the explanatory variable).

a) Produce a scatter plot summarising the relationship between oxygen and runtime and describe the relationship. [3 marks]

b) Estimate the correlation coefficient between oxygen and runtime and interpret it. [2 marks]

c) Regress oxygen on runtime using Stata. Paste the output into your assignment [1 mark]

d) Are the assumptions of simple linear regression satisfied for your regression model? Provide evidence to support your claim. [2 marks]

e) Summarise your regression model in a brief conclusion. [4 marks]