FIT2086 Assignment 2

Due Date: 11:55PM, Monday, 18/9/2023

1 Introduction

There are total of three questions worth $10+10+8=28$ marks in this assignment. This assignment is worth a total of 20% of your final mark, subject to hurdles and any other matters (e.g., late penalties, special consideration, etc.) as specified in the FIT2086 Unit Guide or elsewhere in the FIT2086 Moodle site (including Faculty of I.T. and Monash University policies).

Students are reminded of the Academic Integrity Awareness Training Tutorial Activity and, in particular, of Monash University's policies on academic integrity. In submitting this assignment, you acknowledge your awareness of Monash University's policies on academic integrity and that work is done and submitted in accordance with these policies.

Submission Instructions: Please follow these submission instructions:

1. No files are to be submitted via e-mail. Submissions are to be made via Moodle.
2. You may not use generative A.I. in any capacity when answering this assignment.
3. Please provide a single file containing your report, i.e., your answers to these questions. Provide code/code fragments as required in your report, and make sure the code is written in a fixed width font such as `Courier New` (or a screen shot is taken and inserted – please make sure this is neat and readable), or similar, and is grouped with the question the code is answering. You can submit hand-written answers, but if you do, please make sure they are clear and legible. Do not submit multiple files – all your files should be combined into a single PDF file as required. Please ensure that your assignment answers the questions in the order specified in the assignment. Multiple files and questions out of order make the life of the tutors marking your assignment much more difficult than it needs to be, and may attract penalties, so please ensure your assignment follows these requirements.

Question 1 (10 marks)

In this question we will revisit our analysis of the COVID-19 recovery data that we began in Assignment 1. The file `covid.19.ass2.2023.csv` contains an enlarged version of the New South Wales days-to-recovery data we examined previously with patients with recovery times over four weeks (28 days) removed as these recovery times are unusual and likely represent a sub-population of people more susceptible to the virus. We know from Assignment 1 that the Poisson distribution is not a good fit to the recovery data: instead, for this question we will use a normal distribution as it provides an improved fit to the data due to its increased flexibility, while accepting this assumption is also not necessarily correct; to quote the famous statistician G.E.P.Box: "*all models are wrong – but some are more useful than others*".

Important: you may use R to determine the means and variances of the data, as required, and the R functions `qt()` and `pnorm()` but you must perform all the remaining steps by hand. Please provide appropriate R code fragments and clearly describe all working out.

1. Calculate an estimate of the average number of days to recovery using the provided data. Calculate a 95% confidence interval for this estimate using the t-distribution, and summarise/describe your results appropriately. Show working as required. [**4 marks**]

2. Similar data was collected in 2020 by the Israeli Ministry of Health. While the specific data collected was not available, the summary statistics were provided, and from these I have simulated a dataset of $n = 494$ individuals from the Israeli study. The days to recovery in this group are provided in the file `israeli.covid.19.ass2.2023.csv`. Using the provided data and the approximate method for difference in means with (different) unknown variances presented in Lecture 4, calculate the estimated mean difference in recovery times between the Israeli patients and the patients from NSW, and provide an approximate 95% confidence interval. Summarise/describe your results appropriately. Show working as required. [**3 marks**]

3. It is of interest to determine if there are any differences, at a population level, in recovery times for patients in different countries. Test the hypothesis that the population average time taken to recover for the Israeli cohort is the same as in the NSW cohort. Write down explicitly the hypothesis you are testing, and then calculate a p-value using the approximate hypothesis test for differences in means with (different) unknown variances presented in Lecture 5. What does this p-value suggest about the difference in mean recovery time between the two cohorts of patients? [**3 marks**]

Question 2 (10 marks)

The exponential distribution is a probability distribution for non-negative real numbers. It is often used to model waiting or survival times. The version that we will look at has a probability density function of the form

$$p(y \mid v) = \exp\left(-e^{-v}y - v\right) \tag{1}$$

where $y \in \mathbb{R}_+$, i.e., y can take on the values of non-negative real numbers. In this form it has one parameter: a log-scale parameter v. If a random variable follows an exponential distribution with log-scale v we say that $Y \sim \text{Exp}(v)$. If $Y \sim \text{Exp}(v)$, then $\mathbb{E}[Y] = e^v$ and $\mathbb{V}[Y] = e^{2v}$.

1. Produce a plot of the exponential probability density function (1) for the values $y \in (0, 10)$, for $v = 1$, $v = 0.5$ and $v = 2$. Ensure the graph is readable, the axis are labeled appropriately and a legend is included. **[2 marks]**

2. Imagine we are given a sample of n observations $\mathbf{y} = (y_1, \ldots, y_n)$. Write down the joint probability of this sample of data, under the assumption that it came from an exponential distribution with log-scale parameter v (i.e., write down the likelihood of this data). Make sure to simplify your expression, and provide working. *(hint: remember that these samples are independent and identically distributed.)* **[2 marks]**

3. Take the negative logarithm of your likelihood expression and write down the negative log-likelihood of the data $\mathbf{y}$ under the exponential model with log-scale v. Simplify this expression. **[1 mark]**

4. Derive the maximum likelihood estimator $\hat{v}$ for v. That is, find the value of v that minimises the negative log-likelihood. You must provide working. **[2 marks]**

5. Determine the approximate bias and variance of the maximum likelihood estimator $\hat{v}$ of v for the exponential distribution. *(hints: utilise techniques from Lecture 2, Slide 27 and the mean and variance of the sample mean)* **[3 marks]**

Question 3 (8 marks)

It is frequent in nature that animals express certain asymmetries in their behaviour patterns. It has been suggested that this might be nature's way of "breaking gridlocks" that might occur if we were to act purely rationally (for example, why does a beetle decide to move one way over another when put in a featureless bowl?). An interesting observational study, undertaken by a European researcher in 2003 examined the head tilting preferences of humans when kissing.

The data was collected by observing kissing couples of age ranging from 13 to 70 in public places (mostly airports and train stations) in the United States, Germany and Turkey. The observational data found that of 124 kissing pairs, 80 turned their heads to the right and 44 turned their heads to the left.

You must analyse this data to see if there is an inbuilt preference in humans for the direction of head tilt when kissing. Provide working, reasoning or explanations and R commands that you have used, as appropriate.

1. Calculate an estimate of the preference for humans turning their heads to the right when kissing using the above data, and provide an 95% confidence interval for this estimate using the techniques discussed in Lecture 4. Summarise/describe your results appropriately. **[3 marks]**

2. Test the hypothesis that there is no preference in humans for tilting their head to one particular side when kissing. Write down explicitly the hypothesis you are testing, and then calculate a p-value using the approach for testing a Bernoulli population discussed in Lecture 5. What does this p-value suggest? **[2 marks]**

3. Using R, calculate an exact p-value to test the above hypothesis. What does this p-value suggest? Please provide the appropriate R command that you used to calculate your p-value. **[1 mark]**

4. It is entirely possible that any preference for head turning to the right/left could be simply a product of right/left-handedness. To test this we obtain handedness of a sample of different people. It was found that 83 people were right-handed and 17 were left handed. Using the hypothesis testing procedure for testing two Bernoulli populations from Lecture 5, test the hypothesis that the rate of right-handedness in the population is the same as the preference for turning heads to the right when kissing this data. Summarise your findings. What does the p-value suggest? **[2 marks]**