FIT2086 Assignment 2

Due Date: 11:55PM, Friday, 16/9/2022

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. 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 you assignment follows these requirements.

Question 1 (10 marks)

In this question we will analyse some topical and relevant data: daily reported case numbers of people in Victoria, Australia infected with the novel coronavirus (Covid-19). The data we will use was obtained from the the Victorian government public health website. In particular, we will analyse some daily case numbers for the month of August. It is obviously important for authorities to use data such as this to determine trends in case numbers and make predictions about future loads on the healthcare system. We will start with the daily reported case numbers for the first seven days of August in the file `daily.covid.aug1to7.csv`.

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 all working out.

1. Calculate an estimate of the average number of daily reported cases 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. The file `daily.covid.aug8to14.csv` contains data on daily reported case numbers for the second 7-day period in August. 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 reported daily Covid-19 cases between the first 7 day block of August and the second 7 day block in August, and provide an approximate 95% confidence interval. Summarise/describe your results appropriately. Show working as required. **[3 marks]**

3. It is potentially of interest to see if the daily reported case numbers are changing over time. Test the hypothesis that the population average daily reported case numbers between the two seven-day blocks is the same. 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 average reported daily case numbers between the two seven-day blocks? **[3 marks]**

Question 2 (10 marks)

The negative binomial distribution is a probability distribution for non-negative integers. It models the number of heads observed in a sequence of coin tosses until the r-th tail is observed. As such it is used widely throughout data science to model the number of times until some specific binary event occurs, i.e, the number of years between multiple natural disasters, etc. The version that we will look at has a probability mass function of the form

$$p(y \mid v, r) = \binom{y + r - 1}{y} r^r \left(e^v + r\right)^{-r-y} e^{y\,v} \tag{1}$$

where $y \in \mathbb{Z}_+$, i.e., y can take on the values of non-negative integers. In this form it has two parameters: v, the log-mean of the distribution, and r, the number of tails we are waiting to observe. Often r is not treated as a learnable parameter, but rather is set by the user depending on the context. If a random variable follows a negative binomial distribution with log-mean v we say that $Y \sim \mathrm{NB}(v, r)$. If $Y \sim \mathrm{NB}(v, r)$, then $\mathbb{E}[Y] = e^v$ and $\mathbb{V}[Y] = e^v(e^v + r)/r$.

1. Produce a plot of the negative binomial probability mass function (1) for the values $y \in \{0, 1, \ldots, 25\}$, for ($v = 0$, $r = 1$), ($v = 1$, $r = 2$) and ($v = 1.5$, $r = 2$). Ensure that the graph is readable, the axis are labelled appropriately and a legend is included *(hint: the* `choose()` *function in R may be useful).* **[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 a negative binomial distribution with parameters v and r (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 negative binomial model with parameters v and r. Simplify this expression. **[1 mark]**

4. Derive the maximum likelihood estimator $\hat{v}$ for v, under the assumption that r is fixed; that is, find the value of v that minimises the negative log-likelihood, treating r as a fixed quantity. You must provide working. **[2 marks]**

5. Determine expressions for the approximate bias and variance of the maximum likelihood estimator $\hat{v}$ of v for the negative binomial distribution, under the assumption that r is fixed. *(hints: utilise techniques from Lecture 2, Slide 22 and the mean/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 (think: why does a beetle decide to move one way over another when put in a featureless bowl?).

An interesting study regarding preferences was undertaken by Irish researchers in 2006. In the experiment, 240 volunteer students from Stanmillis University College in Belfast were asked to stand directly in front of a symmetrical doll's face and asked to kiss the doll on the cheek or lips; researchers then recorded whether the student tilted their head to the right or left when kissing the doll. Of the 240 students, 176 turned their head to the right and 64 turned their head 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 approximate 95% confidence interval for this estimate. 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 approximate 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 the handedness of the 240 volunteers was also recorded. It was found that 210 of the participants were right-handed and 30 were left handed. Using the approximate hypothesis testing procedure for testing two Bernoulli populations from Lecture 5, test the hypothesis that the rate of right-handedness in the population from which the participants was drawn is the same as the preference for turning heads to the right when kissing. Summarise your findings. What does the p-value suggest? **[2 marks]**