Question 1 (11 marks)

The fuel efficiency of cars is usually measured in the number of kilometers (on average) that a car can travel on one litre of fuel, under "typical" conditions. Higher fuel efficiency is obviously desirable. The file fuel.efficiency.csv contains records on a subset of actual vehicles measured for fuel efficiency by the US government in the period 2017-2020. The data has fuel efficiency recordings on a number of vehicles along with information indicating whether they are either all-wheel-drive (coded as A) or part-time four-wheel-drive (coded as P). Please use this file to answer the following questions.

Important: you may use R to determine the means and variances of the data, as required, and the R functions pt() 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 fuel efficiency of vehicles that are all-wheel drive. 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. An obvious and important question is: is there a difference in fuel efficiency between all-wheeldrive vehicles and part-time four-wheel drive vehicles? Using the provided data and the approximate method for difference in means with unknown variances presented in Lecture 4, calculate the estimated mean difference in fuel efficiency between all-wheel-drive vehicles and part-time four-wheel-drive vehicles, and a 95% confidence interval for this difference. Summarise/describe your results appropriately. Show working as required. **[3 marks]**

3. Given that in an all-wheel-drive vehicle all four wheels are continuously powered, while in a part-time four-wheel-drive vehicle all wheels are powered only at certain times, it seems plausible that all-wheel-drive cars may have worse fuel efficiency. Using the provided data, test the hypothesis that all-wheel-drives are less efficient than part-time four-wheel-drive vehicles. Write down explicitly the hypothesis you are testing, and then calculate a *p*-value using the approximate hypothesis test for differences in means with unknown variances presented in Lecture 5. What does this *p*-value suggest about the difference between vehicles with all-wheel-drive and part-time four-wheel-drive transmissions? Show working as required. **[3 marks]**

4. Can you identify any possible problems with your conclusions based on the available data? Could there be an alternative explanation for the results you obtained other than their difference in drive-systems (all-wheel-drive *vs* part-time four-wheel-drive)? **[1 mark]**

Question 2 (10 marks)

The geometric distribution is a probability distribution for non-negative integers. It models the number of tails observed in a sequence of (weighted) coin tosses until the first head 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 major natural disasters, etc. The version that we will look at has a probability mass function of the form

$$p(y \mid L) = e_L + 1_{-y-1}\, e_{yL} \qquad (1)$$

where $y \in \mathbb{Z}_+$, i.e., y can take on the values of non-negative integers. In this form it has one parameter: L, the log-odds of seeing a failure (tail) when the coin is tossed. If a random variable follows a geometric distribution with log-odds L we say that $Y \sim \text{Exp}(L)$. If $Y \sim \text{Exp}(L)$, then $\mathbb{E}[Y] = e^L$ and $\mathbb{V}[Y] = e^L(e^L + 1)$.

1. Produce a plot of the geometric probability mass function (1) for the values $y \in \{0,1,...,20\}$, for $L = 0$, $L = 1$ and $L = 2$. Ensure that the graph is readable, the axis are labelled appropriately and a legend is included. **[2 marks]**

2. Imagine we are given a sample of n observations $\mathbf{y} = (y_1,...,y_n)$. Write down the joint probability of this sample of data, under the assumption that it came from a geometric distribution with log-odds parameter L (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 loglikelihood of the data $\mathbf{y}$ under the geometric model with log-odds L. Simplify this expression. **[1 mark]**

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

5. Determine the approximate bias and variance of the maximum likelihood estimator $\hat{L}$ of L for the geometric distribution. *(hints: utilise techniques from Lecture 2, Slide 21 and the mean/variance of the sample mean)* **[3 marks]**

Question 3 (8 marks)

This question is a bit light hearted in nature. It was believed for a long time by medical practitioners that the full moon influenced the expression of medical conditions including fevers, rheumatism, epilepsy and bipolar disorder – in fact, the antiquated term "lunatic" derives from the word lunar, i.e., of the moon. In the late 1990's a (tongue in cheek) study was undertaken to test if the full moon induced dogs to become more aggressive, with a resulting increased likelihood of biting people. The data collected was the daily number of admissions to hospital of people being bitten by dogs from 13th of June, 1997 through to 30th of June, 1998[1]. The average number of dog-bite admissions per day was 3.6. I have converted the data into binary form by denoting a day with less than four dog-bite admissions as a "below average day", and a day with four or more dog-bite admissions as an "above average day".

From the large, complete data set the proportion of non full-moon days that experienced an above average number of dogbite admissions was found to be 0.53. You can treat this as exactly known. There was data available on 26 days that fell on a full moon; of these, 11 had an above average number of dogbite admissions and 15 had a below average number of dogbite admissions.

You must analyse this data to see if the phase of the moon really does have an effect on the aggressiveness of dogs! Provide working, reasoning or explanations and R commands that you have used, as appropriate.

[1] Data source is taken from the Australian Institute of Health and Welfare Database of Australian Hospital Statistics.

1. Calculate an estimate of the probability of a full moon day experiencing an above average number of dog bite admissions 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 difference in the probability of experiencing an above average number of dog bite admissions between full moon and non-full moon days. 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. A researcher suggests that perhaps another way to test whether the phase of the moon has an effect on the aggressiveness of dogs is to compare different phases. In the collected data there were 26 days that fell on a new moon, and of these 20 experienced an above average number of dogbite admissions and 6 had a below average number of dogbite admissions. Using the approximate hypothesis testing procedure for testing two Bernoulli populations from Lecture 5, test the hypothesis that the probability of experiencing an above average number of dogbite admissions does not differ between days falling on the new moon and the full moon. Summarise your findings. What does the *p*-value suggest? **[2 marks]**