CFAS440 Coursework Statistical Methods

Submission date: Monday 21st November 2016, 5pm

- You should answer all questions. The total number of marks available is 65.
- Solutions can be either typed using Latex, or a word processor, or hand written. If hand written please write *clearly*, as I can only mark what I can read. You should submit one hard copy in the homework box and also an electronic version (*i.e.* scanned if handwritten) *via* Moodle.
- Data for the project can be found in the file CWData.Rdata.
- Full marks will only be awarded for correct solutions which are presented in a *clear* manner with appropriate explanations (see next point).
- Wherever appropriate you should state your hypotheses, name the statistical test used and explain why it is appropriate, explain what assumptions have been made, give the test statistic and p-value and/or confidence interval and explain your conclusions in relation to the original problem.
- Any graphs should be interpreted with a sentence or two to summarise key findings. Axes should have appropriate labels, including units of measurement where necessary.
- Try to use the examples in the lecture slides as a guide to presentation.

Q1 Charles Darwin conducted an experiment into the relative growths of cross- and self-fertilised seedlings. In his experiments pairs of seedlings were grown in near identical conditions. In each pair, one seedling had been self-fertilised and the other cross-fertilised. The final heights (in inches) of the plants can be found in the data set `seedlings`.

(a) Obtain sample mean heights for both the cross-fertilised and self-fertilised plants. What are the standard errors of these means? [3]

(b) Obtain a 95% confidence interval for the mean height of the cross-fertilised plants. You should use your answer to part (a) in order to do this. [3]

(c) If you carried out the experiment 500 times, and calculated the 95% confidence interval for the mean height of the cross-fertilised plants each time, how many of the confidence intervals would you expect to contain the population mean? [1]

(d) Using your confidence interval, test whether there is evidence that the mean height is different to 19 inches. [3]

Q2 This question also uses the results of Darwin's experiment into plant growth, but focuses on the differences between the heights of the cross- and self-fertilised plants.

(a) Draw a scatter plot to compare the final heights of the two types of plants. What is the correlation between the two sets of heights? Interpret your results. [3]

(b) Write down null and alternative hypotheses to test whether the final heights of the cross-fertilised plants are greater than the self-fertilised plants. [2]

(c) Carry out an appropriate parametric test on the hypotheses in part (b). You should test at the 5% level. Would you draw the same conclusion at the 1% level? [4]

(d) If you wanted instead to carry out a non-parametric test in part (c), which one would you use? [1]

Q3 The Exponential distribution, which is used to model positive random variables, has a single parameter λ. The expectation and variance for an exponential random variable Y ($Y > 0$) are

$$\mathbb{E}[Y] = \frac{1}{\lambda}, \quad \text{Var}(Y) = \frac{1}{\lambda^2}.$$

(a) What is the method of moments estimate for λ? [2]

The exponential distribution can be used to model the size of non-zero rainfalls. The `rainfall` dataset which can be found in the file CWdata.Rdata contains 285 non-zero hourly rainfall totals (in mm), measured at a single location.

(b) Assume that these observations are an i.i.d sample taken from an Exponential(λ) distribution. Calculate the sample mean for this data, and hence obtain the method of moments estimate of λ. [3]

(c) Using your answer to part (b), what is the probability that the rainfall total in a given hour exceeds $35mm$, $\Pr[Y > 35]$? [2]

(d) Plot a histogram of your data, and overlay this with the density of the Exponential distribution, taking λ to be the value estimated in part (b). Why is the exponential distribution not appropriate? *Hints.* You might want to use the function `dexp` to draw the density. To get an appropriate range of `x` values on which to plot the density, consider both the range of your data and the sampling space of the exponential distribution. [3]

Q4 A more appropriate distribution for modelling non-zero rainfall totals is the Gamma(α, β) distribution. Here $\alpha > 0$ is the shape parameter and $\beta > 0$ is the rate parameter. The expectation and variance for a gamma random variable Y ($Y > 0$) are

$$\mathbb{E}[Y] = \frac{\alpha}{\beta} \quad \text{and} \quad \text{Var}(Y) = \frac{\alpha}{\beta^2}$$

(a) Give expressions for the method of moments estimators for α and β. [5]

(b) Using your answer to part (a), and by calculating the sample mean and variance for the `rainfall` data used in question 2, obtain estimates for α and β under the Gamma model. [3]

(c) Plot the density of your fitted model, using the function `dgamma`. Why is this a more appropriate model for the rainfall data than the exponential model used in the previous question? [2]

(d) What is the estimate of $\Pr[Y > 35]$ under this model? [1]

(e) Using bootstrap methods obtain the sampling distribution for $\Pr[Y > 35]$ under this model, and consequently give an estimate of the 95% confidence interval for this probability. [4]

Q5 A group of trees, consisting of two different species A and B, were cleared of ants using insecticide. A colony of ants was then released in the vicinity of the trees and, after a week, each tree was investigated to discover whether or not it had been colonised. The results are given below and are contained in the table `ants`.

	Invaded	Not invaded
Species A	2	13
Species B	10	3

(a) Calculate the row, column and overall totals for the above data. [2]

(b) Carry out a χ^2 test to decide whether or not ants prefer one species of tree over the other. [4]

(c) What is Fisher's exact test, and when is it useful? [2]

(d) Carry out Fisher's exact test on the ant data. Do your conclusions change? [2]

Q6 In the species of ants investigated in Q5, the worker ants are believed to fall into 5 size categories. A previous study claimed that the proportions of workers in each category were 20%, 30%, 20%, 25% and 5% (small to large). For a random sample of ants from the colony used in the tree colonisation study, the numbers in each category were found to be 8, 14, 8, 6, 4 (small to large). The actual lengths (in mm) of each of the ants in the sample, according to group, can be found in `antLengths`. Groups are ordered from 1 (smallest) to 5 (largest).

(a) Carry out an appropriate test to see whether or not the ant colony in the tree colonisation experiment is consistent with the findings of the original study, in terms of the proportions of workers in each of the size categories. [4]

(b) Produce an appropriate plot of the ant lengths according to group. [2]

(c) What test would be appropriate to assess whether or not there is evidence that the mean ant lengths differ between the five groups? Carry out this test at the 5% level. What assumptions have you made about the data in order to carry out this test? [4]