MAST90044 Thinking and Reasoning with Data
Semester 1 2019
Assignment 2
Due: 8 AM, Tue 5 May

- Assignments are to be submitted via LMS only.
- Please label your assignment with the following information:
  - your name;
  - your student number;
  - the day of your lab class.
- Late assignments will only be accepted under exceptional circumstances and must be discussed with Dr Julia Polak. If it is a medical issue, a medical certificate will be required. A late penalty may be imposed.
- This assignment is worth 15% of the marks in this subject, and covers the work done in weeks 4 to 6.
- The total number of marks for this assignment is 52. The number of marks given for each question may be fine-tuned.
- Tutors will not help you directly with assignment questions. However, they may give some help with R.
- Worded answers should not be more than 3 sentences, unless otherwise stated. Anything greater will not be assessed.
- R commands are to be included that are directly relevant, as well as R output. Marks will be deducted for irrelevant and unnecessary commands and output.
- Questions and parts of questions should be clearly marked, fontsize and general formatting clear and readable. Marks will be deducted for poor formatting.
- Assignments should be restricted to no more than six (two-sided) A4 pages. These will not be marked otherwise.
- Solutions to the assignment questions will be made available later.

Q1 400 random water samples were collected from an aquifer. A total of 183 of these samples contained a contaminant (a pathogenic bacterium). Historically, the average rate of contaminated samples was 42%

(a) Calculate a p-value using an exact procedure and an approximate procedure for the null hypothesis of no change, and calculate 95% confidence intervals for the proportion of contaminated samples. Is there evidence of an increase in the frequency of contaminated samples from this aquifer?

(b) Discuss briefly the potential for and the implications of Type I (in no more than 2 sentences) and Type II (in no more than 2 sentences) errors in this situation.

[6 + 4 = 10 marks]

Q2 Periodic measurements of salinity and water flow were taken in North Carolina's Pamlico Sound, resulting in the following data (x = water flow, y = salinity):

x	23	24	26	25	30	24	23	22	22	24	25	22	22	22	24
y	7.6	7.7	4.3	5.9	5.0	6.5	8.3	8.2	13.2	12.6	10.4	10.8	13.1	12.3	10.4

(a) Read the data into R and produce a suitable graphical summary (with meaningful labels) of the relationship between water flow and salinity.

(b) Write down an appropriate statistical model for examining the relationship, and fit the model in R. Use the regression output to determine the correlation coefficient between x and y.

(c) Examine appropriate diagnostic plots, and comment on anything that is noteworthy or that may challenge the assumptions of the model.

(d) Find a 99% confidence interval for the slope of the line. Comment on the usefulness or otherwise of the estimated slope and intercept.

(e) Find a 95% prediction interval for the salinity when the water flow is 21. Explain its meaning.

[3 + 5 + 5 + 6 + 3 = 22 marks]

Q3 The per diem fecundity (number of eggs laid per female per day for the first 14 days of life, hint - continuous variable) was recorded for 25 females of each of three genetic lines of the fruitfly *Drosophila melanogaster*. The lines labelled R and S were selectively bred for resistance and susceptibility to DDT, respectively, and the line labelled N was a nonselected control strain. The data are in the file `fruitfly.csv` on the LMS. Read it into R.

(a) Use a suitable graphical tool to examine the relationship between fecundity and genetic line, and describe your impressions from the graph.

(b) Formulate a statistical model for analysing the data, and specify the null hypothesis. Define all quantities in your model.

(c) Perform the analysis required to test the null hypothesis, and draw a conclusion from it.

(d) Suppose line N had not been included in the experiment, leaving only lines R and S. Compare the means for lines R and S using a suitable t-test. Compare the confidence interval for the mean difference.

[4 + 6 + 4 + 6 = 20 marks]