MAST90044 Thinking and Reasoning with Data
Semester 1 2019
Assignment 1
Due: 8am, Monday 8 April

Instructions

- Assignments are to be submitted (uploaded) via LMS.
- Please label your assignment with the following information:
  - your name;
  - your student number;
  - your lab class;
  - your tutor's name.
- *You must sign the plagiarism ideclaration.* The link is available on the LMS.
- Your assignment should show all working and reasoning, as marks will be given for method as well as for correct answers. Please spell check your document.
- Paste any R code and output into the appropriate places so that it can be seen easily along with your other work. Graphics from R can be resized within your document; make them smaller as necessary.
- Assignments count for 50% of the assessment in this subject. This one is worth 15%, and covers the work done in weeks 1 to 4.
- Tutors will not help you directly with assignment questions. However, they may give some help with R.
- Solutions to the assignment questions will be made available later.
- When constructing a panel of graphs with multiple plots, it is good to use the R command `par(mfrow = c(nrows,ncols))` where `nrows` is the number of rows and `ncols` the number of columns in the panel. The default is `(1,1)`.

Q.1. In 1994, some second year Agriculture students at this university conducted an experiment at the Mount Derrimut experimental farm. Its purpose was to study the rates of growth of ewes and of their wool, and to study the effectiveness of a lupin diet supplement on increasing ewe and wool growth and quality. The data are in `sheep.csv`, which is available on the LMS. Read it into R.

A total of 30 ewes were used, with 15 randomly selected to receive the lupin supplement. The ewes were initially weighed before the experiment began, and then weighed again after two months on the supplement (or not). The ewes were then shorn and several measurements of their wool growth and quality were made. The variables were:

`diet`	1 = no lupin, 2 = lupin
`initwt`	initial ewe weight (kg)
`wt`	final ewe weight (kg)
`growth`	wool growth (gm/100 sq cm/day)
`length`	fibre length growth (mm/day)
`diam`	fibre diameter (microns)
`prickle`	prickle factor (%fibres > 30 microns)
`yield`	wool yield (% of fleece that is wool)

Ignoring diet:

(a) Summarise the fibre length growth data using summary statistics and two graphical tools. Briefly describe the distribution.

(b) Graphically examine the relationship between wool yield and fibre length growth. Comment on the strength or otherwise of the relationship.

(c) Graphically examine the relationship between wool yield and fibre diameter. Calculate the correlation coefficient between the two variables. Formulate a statistical model to describe the relationship. Graphically fit the model, and use it to estimate the parameters in the model (excluding σ). Briefly describe your findings.

(d) Use two graphical tools to compare the observed distribution of prickle factor with a normal distribution. Briefly comment.

Taking diet into account:

(e) Use two graphical tools to examine the relationship between final ewe weight and diet. Create the labels "no lupin" and "lupin" for diets 1 and 2, and then calculate the mean and standard deviation for each diet. Comment on whether there appears to be an effect of diet on final ewe weight.

(f) Write a statistical model to describe the relationship between final ewe weight and diet. Estimate the parameters in the model. A rough estimate of σ will do. State an assumption of the model.

(g) Calculate the weight gain of each ewe, which is the difference between the initial and final ewe weights. Repeat the steps of part (e) for weight gain (instead of final ewe weight), and examine whether the apparent effect of diet is similar to that in part (e). Briefly describe how you would report on the effect of these two diets on ewe weight.

(h) Type `library(lattice)` in R to ensure that the `xyplot()` function is available. Use `xyplot` to examine the relationship between wool growth and fibre diameter for each diet separately. Divide the data into two "data frames", one for each diet. Calculate the correlation coefficient between wool growth and fibre diameter for each diet separately. Comment on the strength and direction of the relationships.

[4 + 2 + 9 + 3 + 7 + 6 + 7 + 7 = 45 marks]

Q.2. The data in the table below come from a report of a survey which investigated whether snoring was related to various diseases; these particular data relate to the presence or absence of heart disease. Those surveyed were classified according to the amount they snored, on the basis of reports from their spouses. (Reference: Norton PG and Dunn EV (1985). Snoring as a risk factor for disease: an epidemiological survey. *Br Med J* 291:630–632).

		None	Occasional	Nearly every night	Every night
Heart disease	Yes	24	35	21	30
	No	1355	603	192	224

(a) Of people who snore every night, find a point estimate of the population proportion with heart disease. Find a 95% confidence interval for this proportion using each of the three methods covered in the lectures and labs. Comment on the appropriateness or otherwise of the Wald and Agresti-Coull methods here.

(b) Comment on the validity or otherwise of the assumptions made in these calculations.

(c) Find a point and an interval estimate of the difference in heart disease proportions between people who snore at least nearly every night and those who snore occasionally or less. Comment on the claim that snoring is unrelated to presence or absence of heart disease.

(d) Assume now that in the city of Urbanirvana, where no-one snores, the results of this survey precisely represent the population, i.e. that the frequencies in the "None" column of the table scale up exactly when applied to the people of the city. 100 people are randomly selected from Urbanirvana and assessed for heart disease. Define the probability distribution of the random variable X which represents the number of people in the sample with heart disease. Find its expectation (mean) and plot its probability mass function.

(e) Find the probability that

i. Exactly 2 people in the sample have heart disease;

ii. At least 2 people in the sample have heart disease.

[7 + 3 + 5 + 4 + 3 = 22 marks]

Q.3. Wendy and Billy are basketball players. Their coach wants to know who is the better free-throw shooter. They each attempt 100 baskets. Wendy scores 68, Billy scores 58. Can we conclude that their underlying abilities are different?

(a) Calculate and show the mean and standard deviation of the differences between their proportion of successes. Comment on the result.

(b) Calculate BY HAND the 95% confidence interval for the difference (show your working). Draw the 95% CI and briefly state your conclusion in context of the study.

(c) Use R to corroborate your calculations.

[3 + 5 + 2 = 10 marks]