

MATH10282 Introduction to Statistics
Semester 2, 2017/18
Coursework assignment using R

The deadline for submitting this coursework is 4.00pm on Friday 20th April 2018. You should upload your report to the assessment entitled 'R Coursework Submission' on the MATH10282 Blackboard site by this time and date. Please note that a similarity report on your work will be generated by Turnitin to detect plagiarism.

This coursework comprises 10% of the overall marks for the module.

Instructions

- (a) You should prepare your coursework report using Microsoft Word. A PDF report from another program is also acceptable. You can include code and numerical results from R by copying and pasting into your Word document. Comments and discussion of the results should be added as required. You can save any plots created in R, for example as a PDF, and import these into your final report. If you wish to include handwritten section (e.g. with mathematical notation), you may do so by including a scan or photo of the handwritten part in your report.

Include in your report all R commands used to generate results.

- (b) To facilitate anonymous marking, please do not include your name in your report. You should include your Student ID in the title of your submission on Turnitin, and on the first page of your report.
- (c) If you have any queries or problems, please contact me as soon as possible.

Tim Waite, March 2018

Question 1 of 1

The data for this question are contained in the file `incomes.txt`, available on Blackboard. The data consist of the weekly incomes, in pounds (£), of 550 randomly selected individuals from a population.

- (a) Read the data into a data frame called `incomes` in R. [2]
- (b) We are now told by the data collection team that the first 300 rows are incomes from 2014, and the next 250 rows are incomes from 2004. Put this information into a second column in the data frame, called `incomes$year`. [*Hint: you may wish to use the command `rep`.*] [2]
- (c) Draw a box plot for the income distribution in 2014 and another box plot for the income distribution in 2004. To facilitate comparison, put the two box plots on the same axes. Comment on the main features of the plots. Do you think that the distributions are symmetric? [4]
- (d) Put the 2014 income values into a variable called `inc14`. [2]

Question 1 continued overleaf, please turn over

- (e) Let x_1, \dots, x_{300} denote the 2014 data. Suggest a suitable transformation, g , that may enable a normal distribution, $N(\mu_Y, \sigma_Y^2)$, to be fitted to the transformed income data, $y_i = g(x_i)$. Show that suitable estimates of the parameters are $\hat{\mu}_Y = 6.2541$, $\hat{\sigma}_Y^2 = 0.2428$. [3]
- (f) Plot a density histogram of the transformed income values from 2014, and superimpose the p.d.f. of the fitted normal density on the same plot. Comment on the goodness of fit. [3]
- (g) Use the fitted model to estimate the probability that a randomly selected individual in 2014 has a weekly income between £1000 and £1200. Compare this with the corresponding empirical probability. [4]

[Total 20 marks]