7CCMMS61T Statistics for Data Analysis Assignment 1

October 2016

Deadline 5.00am 18 November 2016

Question 1

The file `protein.csv` contains data from several European countries in the 1980s on consumption of different categories of food.

(a) For each variable, calculate appropriate summary statistics to show the level and spread of the data (one statistic for each is enough).

(b) For each variable, plot the data in a suitable way to illustrate the level and the spread.

(c) Calculate a summary statistic to show the association of the consumption of fruit and vegetables with each of the other food categories.

(d) Show a plot illustrating the association of the consumption of fruit and vegetables with each of the other food categories.

Question 2

The file `FlightDelays.csv` contains data on flight delays from airports in the USA for one month in 2015. We will consider the following variables:

- `schedtime` the scheduled time of departure
- `carrier` the airline company
- `deptime` the actual time of departure
- `origin` the airport from which the flight departed
- `weather` 0 = normal; 1 = severe
- `delay` whether the flight departed on time or late

You can ignore the other columns for this exercise.

(a) State the scaling of each of the above variables.

(b) First consider the variable `delay`. Recode the categories as 0 and 1 and plot the recoded variable against each of the other variables. Describe any interesting patterns.

(c) Create a new variable measuring the delay, i.e. the difference between the actual departure time and the scheduled departure time. Plot this against any of the other variables which might show interesting associations. Describe the pattern in each case.

(d) Which of the above two sets of plots is more informative and why?

(e) For any one variable which seems to be associated with the delay, calculate a summary statistic which will describe the strength of the association.