FIT3152 Data analytics. Tutorial 01: Introduction to R/Review of basic statistics

Note, much of the data for the following questions has been sourced from http://www.statsci.org/datasets.html and links within.

1. Using the data sets provided as csv files and the lecture notes, try and reproduce all of the statistics and graphics from Lecture 1.

   Files are: {InvestA, InvestB, Toothbrush, Workers & Concrete}.csv

2. The following data records the length of rivers in the South Island of New Zealand. The lengths are given in kilometres. Data is grouped depending on where it flows into. Source: http://www.statsci.org/data/oz/nzrivers.html

   Pacific Ocean:
   209, 48, 169, 138, 64, 97, 161, 95, 145, 90, 121, 80, 56, 64, 209, 64, 72, 288, 322.

   Tasman Sea:
   76, 64, 68, 64, 37, 32, 32, 51, 56, 40, 64, 56, 80, 121, 177, 56, 80, 35, 72, 72, 108, 48.

   (a) Calculate the summary stats for each group of rivers. Draw a boxplot.
   (b) Test the hypothesis that rivers flowing into the Tasman Sea are shorter on average than those flowing into the Pacific Ocean. Use a significance of 1%

3. When anthropologists analyze human skeletal remains, an important piece of information is living stature. Since skeletons are commonly based on statistical methods that utilize measurements on small bones, the following data was presented in a paper in the American Journal of Physical Anthropology to validate one such method. Variables are: MetaCarp – Metacarpal bone length in cm, Stature (Height of skeleton) in cm. Source: http://www.statsci.org/data/general/stature.html

MetaCarp	Stature
45	171
51	178
39	157
41	163
48	172
49	183
46	173
43	175
47	173

   Draw a scatterplot of the data with Stature as the vertical axis. Calculate the regression equation predicting Stature from MetaCarp. Comment on the accuracy of the model. Superimpose the line of best fit on your scatterplot.

4. The ocean swell produces spectacular eruptions of water through a hole in the cliff at Kiama, about 120km south of Sydney, known as the Blowhole. The times at which 65 successive eruptions occurred from 1340 hours on 12 July 1998 were observed using a digital watch. Source: http://www.statsci.org/data/oz/kiama.html

   Challenge: download the data into R directly from: http://www.statsci.org/data/oz/kiama.txt (See ATHR page 18) or alternatively use the file: Data: kiama.txt

   Read these data into R, creating a vector named 'kiama'. Calculate the mean, standard deviation. Draw the default histogram. Using help, try and draw an improved histogram of your own design by changing range, class intervals and colour etc.

5. The timber data are for specimens of 50 varieties of timber, for modulus of rigidity, modulus of elasticity and air dried density, arranged in increasing order of magnitude of the density. Source: http://www.statsci.org/data/oz/timber.html

   Read these data into R, creating a data frame named 'timber'. You can use the data file: timber.txt or load directly from: http://www.statsci.org/data/oz/timber.txt

   (a) which variable: elasticity or density is a better predictor of rigidity?

   (b) using your choice of variable calculate the regression equation predicting rigidity, draw a scatterplot of the data, showing the fitted model.

   (c) challenge: calculate the regression equation predicting rigidity as a function of both elasticity and density. Comment on the quality of your model vs the single predictor in (b).

6. Challenge: Using the data: InvestA.csv draw a boxplot. You will need to use the help file to work out the syntax – try ?boxplot as a starting point...

   Using the data: InvestA.csv, now use the 'aggregate' function to calculate the mean of each group. This is similar to the 'tapply' function but returns a data frame. Use help to work out the syntax...

7. Analyse Victorian Retail Turnover: Supermarket and grocery stores for the period Jan 2000 – Dec 2010 using Australian Bureau of Statistics data. You will need to copy the data from the Excel file: 8501.0 Retail Trade, Australia.xls. Try and copy the data directly into R using the clipboard (See ATHR page 18).

   Draw a time series plot of the data and plot the time series decomposition. Comment on the main elements in the time series.