The University of Sydney

**STAT2911 Probability and Statistical Models**

| Semester 1 | Computer Class Week 1 | 2023 |
|---|---|---|

Use RStudio to generate your lab report (you will probably need to go to preferences and change "Weave rnw files using: Sweave" to "knitr").

Your report is due by Friday 5pm Week 1.

For the computer class, note that we are logging on to a Linux network, via a Windows PC. To do this

At the Linux log-in screen, enter your "UNIKEY" LOGIN name as username, e.g. abcd0123. Your **password is your UNIKEY password**.

Press F1 (or right-click on the background) and click rstudio. Use your browser to go to the Canvas site for STAT2911 and look at the R references (under the course outline) and the example we provide under the Lab Assignments section. Do not try to read all of it at the moment, rather use it as a reference while you are attempting the questions.

Please be sure to write your `name and week no.` at the top of your lab report. You should place non-graphics **R** commands between `<<>>=` and `@`, and place graphics commands between `<<fig=TRUE>>=` and `@`. Comments should be added outside these regions. Save your lab report into a file named, say `prac1.rnw`. Now use click `Compile PDF` to get a pdf file. If you are satisfied with the output, print it and hand it in when required.

Please make every effort to use properly formed English sentences in your comments: we critically rely on them in grading your work!

1. A boxplot can provide a useful graphical summary of the data (what other graphical summaries do you know?). It consists of a box with a line drawn somewhere through it, whiskers, and outliers.

   (a) Issue the following command in R (it can be in one line) that creates a boxplot for each of the 7 "datasets" defined by `c(...)`.

   ```
   boxplot(c(1:13), c(1:12,19), c(1:12,19.5), c(1:11,19,19.5), c(2:13,-5),
   c(2:13,-5.5), c(4:10,-5,-5.5,-6,19,19.5,20), c(1:6,6,8:13))
   ```

   (b) Find the quartiles and the IQR of the 8 datasets.

   (c) How is R's default boxplot defined?

2. There is a data.frame `faithful` with two columns `eruptions` and `waiting` giving duration of eruption and waiting time between eruptions for the famous Old Faithful Geyser in Yellowstone National Park, Wyoming, USA.

   (a) Load the data from the base library (`data(faithful)`) and attach it (`attach(faithful)`) so you can use the column vectors by name.

   (b) Obtain summaries (`summary`) for the waiting times.

(c) Set up a graph window to take a $2 \times 2$ array of graphs (`par(mfrow=c(2,2))`) and plot the histograms of eruption duration and waiting times specifying 20 classes for each histogram.

Add the boxplots of `10*eruptions` and `waiting` on the same graph. All three figures should be created in the same block of code, right after `par(mfrow=c(2,2))`.

(d) It is important to see if the similarities in distributional shape is due to association between eruption duration and waiting time. Obtain a graph which will answer this question.

(e) Comment on whether the boxplots are useful for this data.

Consult the TA if you encounter any difficulties.