School of Mathematics

MATH5741M: Statistical Theory and Methods

Continuous Assessment, Part II

Note

This is the second (and last) of two pieces of work which count towards your *continuous assessment* component of this module. Overall, the continuous assessment counts for 20% of the total. This piece of work counts for 12% of the total (i.e. 60% of the continuous assessment).

Work to hand in

Please submit your work **as a PDF file** to the VLE (through "turnitin"). Your report should use no smaller than 12 point fonts, be no less than single-spaced, and no more than 6 pages in total (including any title pages, table of contents, apppendices...).

The deadline is midnight on 12th December, 2018.

Data

We will consider again the vhicle accident data that was used in the first practical. This can be read into R (as before) using the command:

```
xx=read.csv("http://www1.maths.leeds.ac.uk/   #type these two as one line
   ~charles/math5741/DfTaccidents.csv",header=T)
```

The meanings of most of the variable codes can be found in this file:
http://doc.ukdataservice.ac.uk/doc/5683/mrdoc/pdf/5683userguide.pdf.
Other variables not listed are "Urban_or_Rural_Area" (1=Urban, 2=Rural, 3=Unallocated), and Columns 23 and 24, which use: 0="none nearby", –1 = "missing data".

Assignment

You will surely use R to analyze these data, but you should not hand in any R output (except graphs). You should show all the steps (within a "self-contained" document) — almost as if you had used a pocket calculator. Make sure you state any assumptions, as well as your conclusions.

1. Draw a boxplot to compare the number of vehicles involved in an urban area, with the number involved in a rural area. Explain why a transformation of the data may (or may not) be appropriate. Using your transformation (or not) carry out a suitable test to investigate whether the average number of vehicles in an accident differs in urban and rural areas.

2. Using a suitable statistical hypothesis test, investigate whether the frequency of accidents varies by day of the week. Repeat this test using only week-days (excuding Saturday and Sunday).

3. Compute a 95% confidence interval for the expected (mean) number of accidents which occur on a Monday. State your assumptions in computing this interval, and verify whether they are valid.

Some possibly useful R commands

You will need to check the help files for the following:
`chisq.test()` and see examples in help pages. If you save this as an object, the components contain some useful information.
`table()` tabulates fequencies, and cross-tabulation. If you save the output as an object, then you can obtain the column headings with `names`, and the frequencies themselves with `as.numeric`.
`as.matrix()` is useful for saving a table, in case you later want to pool rows and/or columns.
`t.test()` this should be familiar from last practical
`shapiro.test()` see help pages..