# Univariate Data and Modelling – Exercises

## Session 6 – One-way ANOVA

**Exercise 1**

Load the SENIC.DAT dataset from Toledo, we already used this dataset during session 3. This is a historic study (1975 – 1976) on the reduction of hospital-acquired infections by means of infection surveillance and control programs. Each observation is the data of one US hospital. It has the following variables:

```
ID          Identification Number
length      Average length of stay of all patients
age         Average age of all patients
risk        Risk of acquiring infection during stay (percentage)
cult        Number of routine cultures performed on patients without symptoms (times 100)
xray        Number of routine X-rays performed on patients without symptoms (times 100)
beds        Number of beds in hospital
meds        Medical school affiliation - 1 = Yes; 2 = No
reg         Region - 1 = NE; 2 = NC; 3 = S; 4 = W
cen         Average number of patients in hospital per day
nur         Average number of nurses in hospital per day
fac         Available facilities and services at hospital (percent)
```

a) Construct 3 dummy variables X1, X2, and X3, defined as follows:

| reg | X1 | X2 | X3 |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 |
| 4 | 0 | 0 | 0 |

b) Fit a linear model of infection risk against these three dummy variables, but do not include any other variables in the model. The R code for this is
`lm(risk~X1+X2+X3, data = senic.df)`
State the alternatives, give the decision rule and comment your conclusion;

c) Use the mean values obtained in b) to calculate a 95% confidence interval for these means, using the `tsum.test` function (Session 2 – package BSDA) and the values in the table below:

| Region | Mean | Sample s.d. | Sample size | Lower limit | Upper limit |
|---|---|---|---|---|---|
| 1 | … | … | 28 | … | ... |
| 2 | … | … | 32 | … | ... |
| 3 | … | … | 37 | … | … |
| 4 | … | … | 16 | … | … |

Do you think by looking at this table that any means are significant different?

EXTRA: draw line plots of the confidence intervals to investigate differences between the means;

d) Test, by using the `aov` function and with a significance of 0.05% if the mean infection risk ("risk") is the same in the four regions ("reg").

e) Calculate, using the `TukeyHSD` function, the family-wise confidence intervals of the means. Compare the results with the values in c). Look at the differences plot, what can you conclude from this and compare with b).

f) Check the model conditions and assumptions for the model constructed in b):

- Are the groups independent?
- Is there a constant within-group variance?
- Are the within-group residuals normal distributed?
- Are there any influential observations?