Homework 2

Due: Friday February 21 at 2pm

See general homework tips and submit your files via the course website.

The exercises use the **cardata** and **bupa** data sets defined in **HW2Data.sas** in the Homework 2 folder on the course website. The data sets are based on the Car Evaluation data[1] and the Liver Disorders data[2] from the UCI Machine Learning Repository[3]. The raw data is available from those websites and is also attached to this assignment in compass. The processed data sets contain the following variables.

The **cardata** data set contains the original variables plus the following two classification variables:

- **acceptyesno** – binary variable for whether the car is at least acceptable
- **acclevel**– a ranking variable just used for sorting purposes

Note: With this data set, you can use the **order=data** option to **proc freq** to retain the ordinal ordering of acceptability variables.

The **bupa** data set includes the original variables and a categorical **drinkgroup** variable:

- **drinkgroup** – categorization of the half-pint equivalents of alcoholic beverages drunk per day (less than 1 drink, 1 to 2 drinks, 3 to 5 drinks, 6 to 8 drinks, or 9 or more drinks)

Exercise 1

For the Car Evaluation data set, do the following,

a) Construct a frequency table for **safety** and **acceptable**, and comment on any apparent associations in the data. Does there appear to be association between safety and acceptability rating, and if so, how do they appear to be related?

b) Perform and comment on appropriate tests of association, and interpret the results (what does the analysis tell us about associations between safety of a car and the acceptability of the car?).

c) Repeat parts **a** and **b** using the binary **acceptyesno** variable instead of the four-valued **acceptable** variable. In addition to the questions in parts **a** and **b**, comment on similarities or differences in the findings for the four-valued and binary acceptability ratings.

Exercise 2

It might be suspected that vehicles with low safety might be considered less acceptable in general than those with medium of high safety. But what if only medium and high safety vehicles are considered? Is there any noticeable association when only medium and high safety vehicles are considered?

Repeat the analyses in **Exercise 1** ignoring the **safety=low** values and interpret your findings. For the table using **acceptyesno**, also determine whether or not high safety vehicles are more likely to be rated at least acceptable than medium safety vehicles are.

Exercise 3

For the **bupa** data, perform a one-way ANOVA for alanine aminotransferase (**sgpt**) as a function of drinking group.

a) Comment on statistical significance of the model, the amount of variation described by the model and its practical significance, and whether or not the equal variance assumption can be trusted.

b) Identify and comment on any significantly different groups. What can we infer about differences in alanine aminotransferase among the five drinking groups based on this analysis?

Exercise 4

Repeat **Exercise 3** using gamma-glutamyl transpeptidase (**gammagt**) as the response instead. Also comment on whether there are more significant differences in gamma-glutamyl transpeptidase or alanine aminotransferase across the drinking groups.

[1] https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

[2] https://archive.ics.uci.edu/ml/datasets/Liver+Disorders

[3] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.