Assignment 1: Data Tools

Assignment Start: 12:01 AM Oct. 16

Assignment End: 11:59 PM Oct. 27

Explain your code where appropriate.

Reference: UCI Machine Learning Repository: Wine Dataset

The wine data set consists of chemical measurements on 13 constituents found in each of the three types of wines (cultivars) grown in the same region of Italy.

1. Compute the Alcohol means by `Cultivar`. Briefly discuss the mean differences among the cultivars.

```
wine <- read.csv("wine.csv")
# wine <- read_csv("wine.csv")
# Only cultivar and alcohol are needed
cultivar <- as.factor( wine[, "Cultivar"])
alcohol <- wine[, "Alcohol"]
# Split 'alcohol' by 'cultivar' to form a list with each element containing the data for a group.
# alcohol.list <- split(...)
# Look at the function 'sapply'.
# sapply(...)
```

2. Compute the number of observations in each `cultivar`.

```
# See comments for 1. above.
# Put your R code here.
```

3. Create a function to perform a one-way analysis of variance. The input argument `z` should be a list consisting of (possibly) named components, one for each group. The output should be a named list containing components for the between SS (SS_B), the within SS (SS_W), the between degrees of freedom, and the within degrees of freedom.
   Note: $SS_B = \sum_i n_i(\bar{y}_i - \bar{y})^2$ and $SS_W = \sum_i (n_i - 1)s_i^2$ where n_i is the sample size of group i, $\bar{y}_i$ is the mean of group i, and s_i^2 is the variance of group i. These group statistics can easily be computed using `sapply`. For the grand mean, $\bar{y}$, think about using `unlist` on `z`. Let g be the number of groups and $n = \sum_i n_i$ be the total sample size, which can also be computed by `unlisting` `z`. n and g are needed to compute the between and within degrees of freedom.
   Note: The code should be general for any g and n_i.

```
# Look at sapply for summarizing over the elements of a list.
oneway <- function(z){
# Put your R code here.
}
# the output should be something like: list(ssb = ssb, ssw = ssw, ...)
```

4. Create a function to summarize the output in a one-way ANOVA table, including the F test and p-value. The input argument is the output named list in the previous question. The output should be one-way ANOVA table.
   Note: For computing the p-value look at the R function `pf`.

```
# For your output, mimic the tabular output of the builtin 'summary' function applied to the output of
# Look at the function 'printCoefmat' to form a table.
oneway.table <- function(x){
```

```
# Put your R code here.
}
# The last line should be: printCoefmat(table, P-values = TRUE, ...)
# where table has computed values bound together with cbind(...), etc.
```

5. Your functions should be illustrated with the `wine` data set. The data consists of 178 samples measuring alcohol (the outcome variable) divided among three (3) cultivars (the input variable).

```
# Use alcohol.list as an argument to oneway and save output to alcohol.aov
# Then call oneway.table, etc.
```

You should provide brief explanations of the output. Compare the output to that obtained from the standard R functions `aov` and `summary`.