Assignment 1 Total mark: 20

Read the instructions carefully. The assignment will be submitted in the form of a .doc or .docx file. You will give the `R` *code for each question and comment/interpret it. Comments on each question should not exceed 200 words and the total document should not exceed 10 pages. When including a figure, do not forget to add a succinct legend mentioning exercise number, question number and type of plot.*

Exercise 1 *Box-and-wiskers plot of persons of Golub et al. (1999) data* *[Total mark: 2].*

1. *Use* `boxplot(golub100)` *to produce a box-and-whiskers plot for each column (patient). Make a screen shot to save it in a word processor. Describe what you see. Are the medians of similar size? Is the inter quartile range more or less equal. Are there outliers?* *[1pt]*
2. *Compute the mean and medians of the patients. What do you observe?* *[1pt]*

Exercise 2 *Hypothesis testing 1/2* *[Total mark: 5].*

1. *Gene selection. We are interested in the following list of candidate genes for the Golub study (you will use the* `golub100` *data from the practical):*

```
list.gene = c(
  "LYZ Lysozyme", "CTSD Cathepsin D (lysosomal aspartyl protease)",
  "Clone 23721 mRNA sequence", "Neuromedin B mRNA",
  "DHPS Deoxyhypusine synthase",
  "GB DEF = (lambda) DNA for immunoglobin light chain",
  "Leukotriene C4 synthase (LTC4S) gene", "KIAA0102 gene" ,
  "Non-lens beta gamma-crystallin like protein (AIM1) mRNA, partial cds",
  "CD24 signal transducer mRNA and 3' region")
```

   (a) *For each of these genes, perform a two-sample t-test values for which the ALL mean is greater than the AML mean (test with unequal variances). You will formulate the null and alternative hypotheses, the test statistic used (and the distribution of the test statistic, including the number of degrees of freedom), the rejection region and draw your conclusion. (Advice: you can create a vector called* `p.value` *which will store the p-values associated to the tests, and name* `p.value` *using the function* `names`*).* *[2pt]*

   (b) *Report amongst these genes those that have a p-value < 0.01. We will refer to these genes as 'differentially expressed' genes.* *[1pt]*

   (c) *Illustrate these results by plotting these differentially expressed genes using boxplots. Interpret the boxplots.* *[2pt]*

Exercise 3 *Hypothesis testing 2/2* *[Total mark: 4].*

1. *Gene CD33. Use the* `grep()` *function to find the index of the important gene CD33 among the rownames of* `golub100`*. For each test below, formulate the null and alternative hypotheses, the test statistic used (its value and the distribution of the test statistic, including the number of degrees of freedom), the rejection region and draw your conclusion:*

   (a) *Test the normality of the ALL and AML expression values*[1]. *[1pt]*

   (b) *Test for the equality of variances.* *[1pt]*

   (c) *Test for the equality of the means by an appropriate t-test.* *[1pt]*

   (d) *Is the experimental effect strong?* *[1pt]*

Exercise 4 *Clustering and visualisation* *[Total mark: 9].*

1. *Gene selection. We are still interested in the list of the 10 candidate genes from Exercise 2 for the Golub study.*

```
>   load('../Prac/prac-R/Data/Golub.RData')
>   list.gene = c(
+   "LYZ Lysozyme", "CTSD Cathepsin D (lysosomal aspartyl protease)",
+   "Clone 23721 mRNA sequence", "Neuromedin B mRNA",
+   "DHPS Deoxyhypusine synthase",
+   "GB DEF = (lambda) DNA for immunoglobin light chain",
+   "Leukotriene C4 synthase (LTC4S) gene", "KIAA0102 gene" ,
+   "Non-lens beta gamma-crystallin like protein (AIM1) mRNA, partial cds",
+   "CD24 signal transducer mRNA and 3' region")
> data.sub = golub100[list.gene, ]
> # compute Euclidian distance and Ward linkaage
> golub.dist = dist(t(data.sub))
> golub.ward <- hclust(t(golub.dist), method='ward')

> # output the dendrogram
> plclust(golub.ward, labels=gol.factor, ylab="Ward")
```

   Create a new data frame from the `golub100` *data set so that the data set subset contains only these genes of interest.*

2. *Hirerarchical clustering. Output the heatmap, using the $1-$correlation distance and the average linkage. Comment on the clusters.* *[2pt]*

3. *Now, display an heatmap with the Euclidian distance and the Ward linkage. Comment on the differences with the heatmap obtained above.* *[1pt]*

   In the remaining of the exercise, you will transpose the data and check that the number of rows of the data frame is 38 (the number of patients).

4. *Principal Component Analysis. Apply a PCA on the data frame (use the arguments center and scale).* *[1pt]*

   (a) *Will two components be enough to explain most of the variance in the data? (give some numerical figures).* *[1pt]*

[1] For this particular question, only formulate the null and alternative hypotheses, the value of the test statistic, the rejection region and draw your conclusion)

(b) *Output the sample plot on the first two components. The samples (patients) should appear on this plot. Comment.* *[1pt]*

(c) *Comment on the biplot obtained. By plotting the boxplots on some chosen genes, explain the characteristics of the genes of interest with respect to how they are located on the biplot: are they overexpressed / underexpressed in some biological conditions?* *[2pt]*

(d) *Compare the clustering of the genes observed on the PCA biplot to the K-means clustering with* $k = 2$. *[1pt]*