STATISTICS 446/646 - Exam one (Take home)
March 7, 2019

Name ___________________________________

Rules for take home exam: Please read the follwoing rules and confirm by signing the cover page that you have read and abide them while taking your exam.

(1) There are three pages including this cover page. You must turn in this signed cover page with exam; otherwise you will receive zero.

(2) You have exactly 24 hours to solve the exam. ABSOLUTELY NO EXTENSIONS.

(3) Exam must be submitted via eCampus using exam submission link. The data and the exam with the submission link are posted inside the exam submission folder at the left panel of eCampus.

(4) Please save your submission as a PDF, with format "lastname_firstname.pdf". Include all R code.

(5) The exam must be taken completely alone. Showing it or discussing it with anybody is forbidden, including (but not limited to) the other students in the course in current or previous years.

(6) You may use any publicly available material you want, including books, the internet, course materials at eCampus etc. (You are NOT allowed to submit questions to internet discussion groups, though!).

(7) Make an effort to make your submission clear and readable. Severe readability issues may be penalized by grade.

(8) Submit your code separately (or integrated into the solution or at the end of the solution) with comments and explanations. Even if the final result is wrong, the code may allow me to find the bug and award partial credit.

(9) There are 4 questions on the exam and total 50 points.

(10) GOOD LUCK!

On my honor, as an Aggie, I have neither given nor received unauthorized aid on this exam

Student's Signature________________________________

STAT646 - Exam 1

Daniel Osorio - dcosorioh@tamu.edu
Department of Veterinary Integrative Biosciences
Texas A&M University

1. State the Central Dogma of Molecular Biology in your own words. *The central dogma of molecular biology states that DNA produces RNA and RNA make proteins. The first event occurs in a process called transcription, and the second during a process called translation.*

2. The exam 1 data on eCampus come from Affimetrix gene expression microarrays on human cells (of a particular type) under two conditions. The data were obtained from GEO. There are 8 arrays. Cells from 2 donors were used. For each donor, cells of two types (Th17 and Treg) were isolated. Representatives of each cell type were subjected to either a treatment or control condition. Table 1 summarizes the meta information for these data.

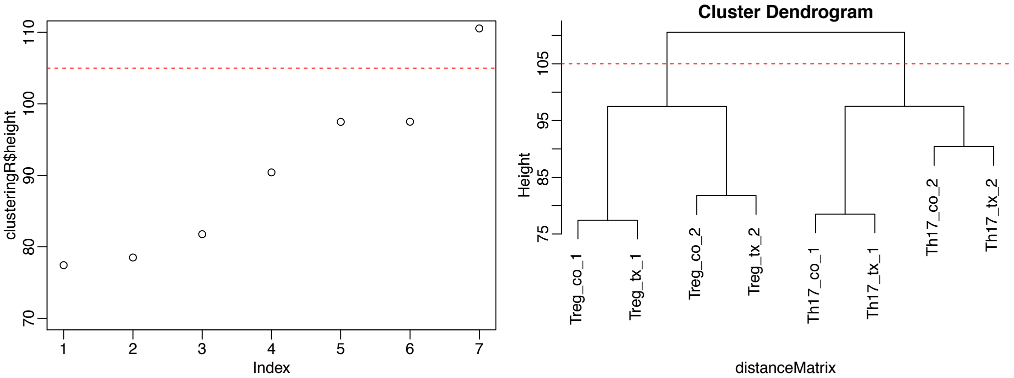
```
> data <- read.csv("exam_1.csv")
```

   (a) Carry out cluster analysis to explore structure among the arrays (columns), using Euclidean distance and complete linkage.

```
> distanceMatrix <- dist(t(data), method = "euclidean")
> clusteringR <- hclust(distanceMatrix, method = "complete")
```

      i. Provide the dendrogram and the plot of merge heights.

```
> par(mfrow=c(1,2), mar=c(2.5,2.5,1,1), mgp = c(1.5,0.5,0))
> plot(clusteringR$height, ylim=c(70,110))
> abline(h = 105, col= "red", lty = 2)
> plot(clusteringR, ylim=c(70,110))
> abline(h = 105, col= "red", lty = 2)
```



      ii. How many clusters would you say there are, and why? *From the height plot, there are two main clusters given by the high separation between the last point and the others. However, from the cluster plot is evident that inside these two there are other two subclusters.*

      iii. How would you explain the apparent clustering? *Main clusters are formed by cell type (Th17 and Treg) and the clusters inside are driven by the donor.*

(b) Carry out principal component analysis to explore structure among the arrays (columns)

```
> PC <- prcomp(t(data))
```

i. Should you standardize or not? Does it matter in this case? In what follows, so we all match answers, do not standardize the columns. *As all the columns seem to have approximately the same variability, it's not required to standardize the columns. See below:*

```
> apply(data,2,var)
Th17_co_1 Th17_tx_1 Treg_co_1 Treg_tx_1
 3.939703  3.926221  3.903509  3.944170

Th17_co_2 Th17_tx_2 Treg_co_2 Treg_tx_2
 3.823271  3.933689  3.959382  3.884836
```

ii. What is the proportion of total variance explained by the first principal component? By the second principal component?

```
> summary(PC)$importance[,1:2]
                            PC1      PC2
Standard deviation     36.74840 29.73041
Proportion of Variance  0.28287  0.18515
Cumulative Proportion   0.28287  0.46802
```

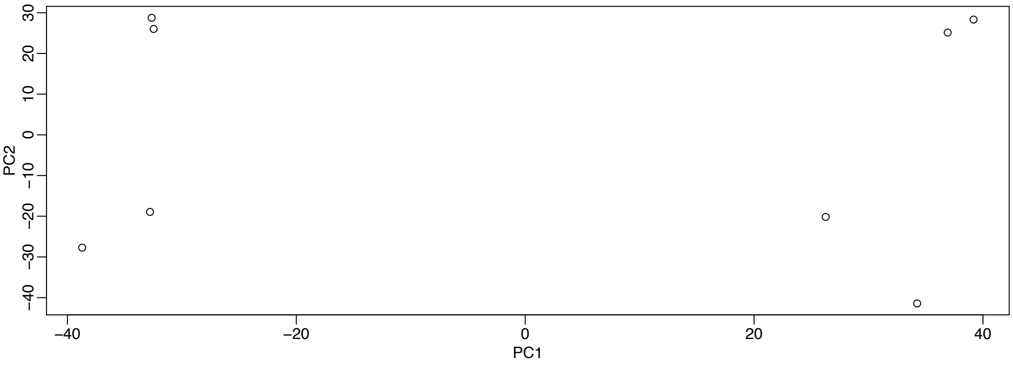
iii. What are the coefficients that define the first principal component's linear combination of the columns (in other words, what is the first eigenvector of the 8×8 sample covariance matrix)? Comment. What do these tell you, if anything, about the data? *The coefficients of this eigenvector split the samples by cell types in the same way as the cluster analysis. All the Th17 have negative values meanwhile all Treg have positive ones.*

```
> PC$x[,1]
Th17_co_1 Th17_tx_1 Treg_co_1 Treg_tx_1
-32.46593 -32.64027  36.92041  39.17638

Th17_co_2 Th17_tx_2 Treg_co_2 Treg_tx_2
-38.71109 -32.78506  26.25808  34.24749
```

iv. Provide a scatterplot of the first two principal components. Comment. Are there any inter- esting patterns, clustering, outliers? If so, how would you explain them? *As well as in cluster analysis there are two main clusters in the PC1 driven by the cell type and other two in the PC2 driven by the donor.*

```
> par(mar=c(2.5,2.5,1,1), mgp = c(1.5,0.5,0))
> plot(PC$x[,1:2])
```

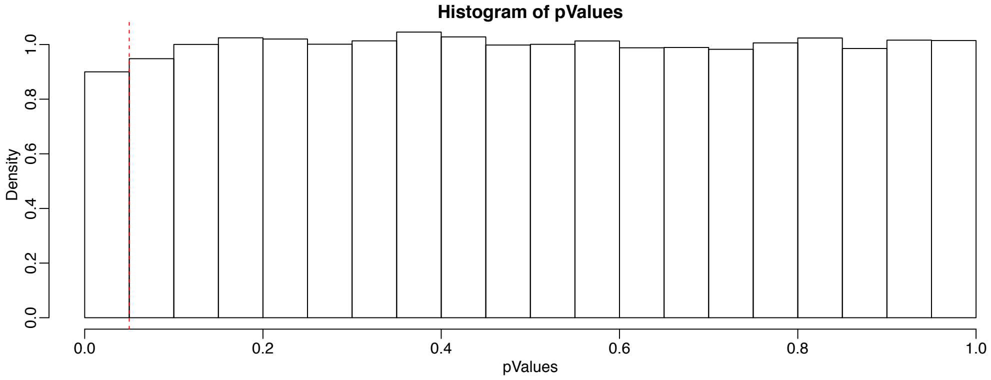


(c) Use two-sample t-tests (assuming unequal variances, the default in R) to test for diffrential expression between the 4 treatment and 4 control arrays

```
> isControl <- grepl("co", colnames(data))
> diffM <- data[,isControl] - data[,!isControl]
> pValues  <- apply(diffM, 1, function(X){
+   t.test(X)$p.value
+ })
```

i. Report a histogram of the p-values. Comment on its shape (does it look as you would expect, are there any worrisome features, do there appear to be many genes for which mean expression differs between control and treatment conditions). *Distribution is uniform. There appear to be less differentially expressed genes than the expected by chance.*

```
> par(mar=c(2.5,2.5,1,1), mgp = c(1.5,0.5,0))
> hist(pValues, probability = TRUE)
> abline(v = 0.05, col = "red", lty = 2)
```



ii. If none of the genes were differentially expressed, what proportion of the p-values would you expect to be <0.05? *Approximately 5%* What proportion of p-values are <0.05? *Around 4.5%* What does this tell you about whether there are many genes for which mean expression differs between control and treatment conditions? *There is not enough evidence supporting a difference between treatment and control samples.*

```
> mean(pValues < 0.05)
[1] 0.04500438
```

iii. Convert the p-values to q-values using p.adjust (use method = "fdr"). How many genes do you call differentially expressed at an estimated FDR of 0.05? *None*

```
> table(p.adjust(pValues) < 0.05)
FALSE
53617
```

(d) The MAX gene is present in multiple species. The versions of the gene in human, mouse are said to be homologous

```
> library(seqinr)
> choosebank("genbank")
```

i. Consider the global alignment of the sequences of this gene in human (Homo sapiens) and mouse (Mus musculus). Use seqinr to download the DNA sequences. You will get multiple sequences for each species. Use third sequence (BC004516.MAX) for human and first sequence (BC138671.MAX) for mouse to answer the following questions. The lengths of the human and mouse sequences will be the same.

```
> hsaMAX <- query(listname="MAX", query="SP=Homo sapiens AND K=MAX")
> hsaMAX <- getSequence(hsaMAX$req[[3]])
> mmusMAX <- query(listname="MAX", query="SP=Mus musculus AND K=MAX")
> mmusMAX <- getSequence(mmusMAX$req[[1]])
```

A. Comparing the human and mouse versions of MAX, what proportion of bases dffer? *Approximately 4%*

```
> mean(hsaMAX != mmusMAX)
[1] 0.03933747
```

B. Compare the human and mouse versions of MAX with respect to GC content of the bases? *Their proportions of GC contents are almost identical around the 50%*

```
> sum(table(hsaMAX)[c("g","c")])/length(hsaMAX)
[1] 0.5403727
> sum(table(mmusMAX)[c("c","g")])/length(mmusMAX)
[1] 0.5341615
```

ii. Consider the global alignment of these two DNA sequences: GAG and GTAG. Let the score for a match be +1, the score for a mismatch be -1, and the score for a gap be -2 . Write down the score matrix for this alignment, and fill in all of the cells. (Note: you do not have to find me any alignment just fill out the score matrix i.e each cell value with direction.)
