Math7340 HW7

Chengbo Gu

Problem 1 (30 points)

For the Golub et al. (1999) data set, use appropriate Wilcoxon two-sample tests to find the genes whose mean expression values are higher in the ALL group than in the AML group.

a) Use FDR adjustments at the 0.05 level. How many genes are expressed higher in the ALL group?

```
data(golub)
gol.fac <- factor(golub.cl, levels=0:1, labels=c("ALL", "AML"))
p.values <- apply(golub, 1, function(x)
  wilcox.test(x~gol.fac, paired = F, alternative = "greater")$p.value )
p.fdr <-p.adjust(p=p.values, method="fdr")
sum(p.fdr < 0.05)
```

```
## [1] 407
```

There are 407 genes that are expressed higher in the ALL group.

b) Find the gene names for the top three genes with smallest p-values. Are they the same three genes with largest difference between the means in the ALL group and the AML group?

```
golub.gnames[,2][order(p.fdr)][1:3]
```

```
## [1] "Macmarcks"
## [2] "VIL2 Villin 2 (ezrin)"
## [3] "TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)"
```

The top three genes with smallest p-values are Macmarcks, VIL2 villin 2 and TCF3 Transcription factor 3.

They are different with the three genes with largest difference between the means in the ALL group and the AML group.

Problem 2 (15 points)

For the Golub et al. (1999) data set, apply the Shapiro-Wilk test of normality to every gene's expression values in the AML group. How many genes do not pass the test at 0.05 level with FDR adjustment? Please submit your R script with the answer.

```
rm(list=ls())
data(golub)
gol.fac <- factor(golub.cl, levels=0:1, labels=c("ALL", "AML"))
p.values <- apply(golub[, gol.fac == "AML"], 1, function(x) shapiro.test(x)$p.value)
p.fdr <-p.adjust(p=p.values, method="fdr")
sum(p.fdr < 0.05)
```

```
## [1] 225
```

There are 225 genes that do not pass the test at 0.05 level with FDR adjustment.

Problem 3 (15 points)

Gene "HOXA9 Homeo box A9" can cause leukemia (Golub et al., 1999). Use appropriate Wilcoxon two-sample tests to test if, for the ALL patients, the gene "HOXA9 Homeo box A9" expresses at the same level as the "CD33" gene. Please submit your R script with the answer.

```
rm(list=ls())
data(golub)
gol.fac <- factor(golub.cl, levels=0:1, labels=c("ALL", "AML"))
index.HOXA9 <- grep("HOXA9 Homeo box A9", golub.gnames[,2])
index.CD33 <- grep("CD33", golub.gnames[,2])


# Signed-ranks Test
data(golub, package='multtest')
wilcox.test (x= golub[index.HOXA9, gol.fac == "ALL"],
             y= golub[index.CD33, gol.fac == "ALL"], paired=T, alternative="two.sided")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  golub[index.HOXA9, gol.fac == "ALL"] and golub[index.CD33, gol.fac == "ALL"]
## V = 62, p-value = 0.01242
## alternative hypothesis: true location shift is not equal to 0
```

Since p-value = 0.01242 is very small, we reject the null hypothesis.

We conclude that the two genes do express differently.

Problem 4 (20 points)

The data set "UCBAdmissions" in R contains admission decisions by gender at six departments of UC Berkeley. For this data set, carry out appropriate test for independence between the admission decision and gender for each of the departments.

```
apply(UCBAdmissions, 3, function(x) fisher.test(x)$p.value)
```

```
##           A           B           C           D           E           F
## 1.669189e-05 6.770899e-01 3.866166e-01 5.994965e-01 3.603964e-01 5.458408e-01
```

The p-value of department A is less than 0.05 while the p-values of other 5 departments are greater than 0.05. Thus, we conclude that admission decision and gender are not independent of department A while admission decision and gender are independent of B, C, D, E and F departments.

Problem 5 (20 points)

There are two random samples $X_1 ... X_n$ and $Y_1 ... Y_m$ with population means μ_X and μ_Y and population variances σ_X^2 and σ_Y^2 . For testing $H_0 : \sigma_X^2 = \sigma_Y^2$ versus $H_A : \sigma_X^2 < \sigma_Y^2$ we can use a permutation test for the statistic $S = \frac{S_X^2}{S_Y^2}$.

Please program this permutation test in R. Use this nonparametric test on the "CD33" gene of the Golub et al. (1999) data set. Test whether the variance in the ALL group is smaller than the variance in the AML group. Please submit your R code with the answer.

```r
rm(list=ls())
data(golub, package='multtest')
gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))
index.CD33 <- grep("CD33", golub.gnames[,2])
data<-golub[index.CD33, ]
n<-length(data)
T.obs<- var(data[gol.fac=="ALL"])/var(data[gol.fac=="AML"])

#Observed statistic
n.perm=2000
T.perm = rep(NA, n.perm)
for(i in 1:n.perm) {
  data.perm = sample(data, n, replace=F) #permute data
  T.perm[i] = var(data.perm[gol.fac=="ALL"])/var(data.perm[gol.fac=="AML"]) #Permuted statistic
}
mean(T.perm<=T.obs) #p-value
```

```
## [1] 0.0365
```

The p-value is 0.0365 that is less than 0.05.

Thus, we reject the null hypothesis and conclude that the variance in the ALL group is smaller than the variance in the AML group.