Solution to Homework module 7

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?

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?

Please submit your R commands together with your answers to each part of the question.

Solutions:

a) The no. of genes expressed higher in ALL group after the FDR adjustment at 0.05 level is 407.

The R code for 1a) is :

```
> # load the data
> data(golub, package='multtest')
> gol.fac<-factor(golub.cl, level=0:1, labels=c("ALL","AML"))
>
> #for Wilcoxon two-sample tests
> wilcox.data=NULL
> for (i in 1:3051){
+   wilcox.data[i]<-wilcox.test (golub[i,] ~gol.fac, paired=F, alternative="g
reater")$p.value
+ }
There were 12 warnings (use warnings() to see them)
>
> # to find the genes expressed higher in ALL group
>
> gene.exp<-wilcox.data<0.05
> sum(gene.exp)
[1] 698
>
> fdr.wilcox<-p.adjust(p=wilcox.data, method="fdr")
> sum(fdr.wilcox<0.05)
[1] 407
```

Solution to Homework module 7

b) The top three genes names with smallest p-value :

```
> non.fdr<- order(wilcox.data, decreasing=FALSE)
> print("Gene names for genes with top 3 smallest p-values before FDR adjustm
ent")
[1] "Gene names for genes with top 3 smallest p-values before FDR adjustment"
> golub.gnames[non.fdr[1:3],2]
[1] "TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors
E12/E47)"
[2] "Macmarcks"
[3] "VIL2 Villin 2 (ezrin)"
```

```
> fdr.AML<- order(fdr.wilcox, decreasing=FALSE)
> print("Gene names for genes with top 3 smallest p-values afte
r FDR adjustment")
[1] "Gene names for genes with top 3 smallest p-values after FD
R adjustment"
> golub.gnames[fdr.AML[1:3],2]
[1] "Macmarcks"
[2] "VIL2 Villin 2 (ezrin)"
[3] "TCF3 Transcription factor 3 (E2A immunoglobulin enhancer b
inding factors E12/E47)"
```

```
> ALL.mean = apply(golub[, gol.fac=="ALL"], 1, mean)
> AML.mean = apply(golub[, gol.fac=="AML"], 1, mean)
> Difference<- ALL.mean - AML.mean
> diff.order<- order(Difference, decreasing=TRUE)
> print("Largest difference")
[1] "Largest difference"
> golub.gnames[diff.order[1:3],2]
[1] "TCL1 gene (T cell leukemia) extracted from H.sapiens mRNA
for Tcell leukemia/lymphoma 1"
[2] "MB-1 gene"
[3] "GB DEF = (lambda) DNA for immunoglobin light chain"
```

```
>
```

⇨ They are not the same three genes with largest difference between the means in the ALL group and the AML group.

Solution to Homework module 7

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.

Solution: For p-values greater 0.05 we do not reject the NULL HYPOTHESIS as

Values follow normal distribuution

The No. of genes that do not pass the test at 0.05 level with FDR adjustment is 225

The R code for the problem is:

```
> # load the data
> data(golub, package='multtest')
> gol.fac<-factor(golub.cl, level=0:1, labels=c("ALL","AML"))
>
> #applying the test
>  shapiro.test<- apply (golub[, gol.fac=="AML"], 1, function(x
) {
+    shapiro.test(x)$p.value })
>
> # get fdr p values
> fdr<- p.adjust(p=shapiro.test, method="fdr")
>
> # calculating and printing the number of genes that failed te
st
>
> print("the genes do not pass the test at 0.05 level with FDR
adjustment")
[1] "the genes do not pass the test at 0.05 level with FDR adju
stment"
> print(sum(fdr<0.05))
[1] 225



>
```

Solution to Homework module 7

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.

Solution:

```
> # load the data
> data(golub, package='multtest')
> gol.fac<-factor(golub.cl, level=0:1, labels=c("ALL","AML"))
>
> # getting the row index of genes
> HOX<- grep("HOXA9 Homeo box A9",golub.gnames[,2])
> print("the row index of HOXA9 Homeo box A9  ")
[1] "the row index of HOXA9 Homeo box A9  "
> print(HOX)
[1] 1391
> CD33<- grep("CD33",golub.gnames[,2])
> print(" the row index of CD33 is")
[1] " the row index of CD33 is"
> print(CD33)
[1] 808
>
>
> # applying the test
>
> wilcox.test<-wilcox.test (x= golub[1391, gol.fac=="ALL"], y= golub[808, gol
.fac=="ALL"], paired=T, alternative="two.sided")
Warning message:
In wilcox.test.default(x = golub[1391, gol.fac == "ALL"], y = golub[808,  :
  cannot compute exact p-value with zeroes
> print(wilcox.test)

        Wilcoxon signed rank test with continuity correction

data:  golub[1391, gol.fac == "ALL"] and golub[808, gol.fac == "ALL"]
V = 62, p-value = 0.01242
alternative hypothesis: true location shift is not equal to 0
```

As p-value 0.01242 we accept the NULL HYPOTHESIS and conclude that the 2 genes do not express at different levels

Solution to Homework module 7

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.

What are your conclusions? Please submit your R script with the answer.

Solutions:

We can conclude that, considering level of significance at 0.05

For department A:

p-value = 5.205e-05 , which is very small, so the NULL HYPOTHESIS of independence can be rejected, which in turns menas that gender and Admissions are dependent

For the other departments (B – E):

p-value = p = 0.7705, 0.4262, 0.6378, 0.3687, 0.6404 respectively.

These values are greater than 0.05.

Thus the NULL HYPOTHESIS of independence cannot be rejected, which means the gender and admissions are probably independent.

The R code for the problem is :

```
#loading the source

source("http://www.bioconductor.org/biocLite.R")

abiocLite()

library(datasets);
```

Solution to Homework module 7

```
> str(UCBAdmissions)
 table [1:2, 1:2, 1:6] 512 313 89 19 353 207 17 8 120 205 ...
 - attr(*, "dimnames")=List of 3
  ..$ Admit : chr [1:2] "Admitted" "Rejected"
  ..$ Gender: chr [1:2] "Male" "Female"
  ..$ Dept  : chr [1:6] "A" "B" "C" "D" ...
>
> Dept <- c("Dept = A","Dept = B", "Dept = C", "Dept = D", "Dept = E", "Dept
= F")
>
> for (i in 1:6 ){
+   print(Dept[i])
+   Dept.Data <- matrix(c(UCBAdmissions[1,1,i], UCBAdmissions[2,1,i], UCBAdmi
ssions[1,2,i],
+                         UCBAdmissions[2,2,i]), nrow=2, dimnames=list("Admit"
=c("Admitted","Rejected"),
+                                                                      "Gender
"=c("Male","Femeale")))
+ # applying the chi-square test and fisher test and printing the data
+   print(Dept.Data)
+   print(chisq.test(Dept.Data))
+   print(fisher.test(Dept.Data))
+ }
```

Solution to Homework module 7

Problem 5 (20 points)

There are two random samples $X_1 \ldots X_n$ and $Y_1 \ldots Y_m$ with population means μ_X and μ_Y and population variances σ^2_X and σ^2_Y. For testing H_0: $\sigma^2_X = \sigma^2_Y$ versus H_A: $\sigma^2_X < \sigma^2_Y$, 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.

Solutions:

As the p-value is 0.0355

We conclude that the variance of ALL is less that variance of AML i.e.

Var(ALL) < Var(AML) and thus we accept the ALTERNATE HYPOTHESIS .

The R code for the problem is :

```
> library(gtools)
> #load the data
> data(golub, package = "multtest")
> gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML")
)
> #find the row index
> gene <- grep("gene_name",golub.gnames[,2])
>
> data <- golub[gene,]
> n <- length(data)
>
> T.obs <- var(data[gol.fac=="ALL"]) / var(data[gol.fac=="AML"]
)
> #observe statistic
> n.perm = 2000
> T.perm = NULL
> for(i in 1:n.perm) {
+   data.perm = sample(data, n, replace=F)
+   T.perm[i] = var(data.perm[gol.fac=="ALL"]) / var(data.perm[
gol.fac=="AML"])
+ }
>
> mean(T.perm <= T.obs)
[1] NA
>
> # applying the above formula to CD33 gene
```

Solution to Homework module 7

```
> # loading data
> data(golub, package = "multtest")
> gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML")
)
>
> # finding row index
> CD33 <- grep("CD33",golub.gnames[,2])
>
> data <- golub[CD33,]
> n <- length(data)
> # observe statistic
> T.obs <- var(data[gol.fac=="ALL"]) / var(data[gol.fac=="AML"]
)
>
> n.perm = 2000
> T.perm = NULL
> for(i in 1:n.perm) {
+   data.perm = sample(data, n, replace=F)
+   T.perm[i] = var(data.perm[gol.fac=="ALL"]) / var(data.perm[
gol.fac=="AML"])
+ }
> # p-value
> mean(T.perm <= T.obs)
[1] 0.0355



>
```