**Module 7 Home Work**

**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.

Answer:

There are 698 genes whose mean expression values are higher in the ALL group than

in AML group.

(a)

data(golub, package = "multtest")

gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))

# Wilcox 2-sample test:

wcox<- NULL

for(i in 1:3051){

wcox[i] = wilcox.test (golub[i,] ~gol.fac, paired=F, alternative="greater")$p.value

}

wcox.genes <- wcox<.05

sum(wcox.genes)

wcox.fdr<- p.adjust(p=wcox, method="fdr")

sum(wcox.fdr < .05)

There are 407 genes expressed higher in the ALL group FDR adjustments.

(b)

**#before Fdr Adjustment**:

data(golub, package = "multtest")

gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))

AML.before<- order(wcox, decreasing=FALSE)

golub.gnames[AML.before[1:3],2]

Output:

[1] "TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)"

[2] "Macmarcks"

[3] "VIL2 Villin 2 (ezrin)"

**#after Fdr adjustment**

data(golub, package = "multtest")

gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))

AML.after <- order(wcox.fdr, decreasing=FALSE)

golub.gnames[AML.after[1:3],2]

Output:

[1] "Macmarcks"

[2] "VIL2 Villin 2 (ezrin)"

[3] "TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)"

**#Large difference**

data(golub, package = "multtest")

gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))

meanALL = apply (golub[, gol.fac=="ALL"], 1, mean)

meanAML = apply (golub[, gol.fac=="AML"], 1, mean)

diff.mean = meanALL - meanAML

diff <- order(diff.mean, decreasing=TRUE)

golub.gnames[diff[1:3],2]

Output:

[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"

**Conclusion:** From the above output its quite evident that they are not the same three genes with largest difference between means in ALL group and 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.

Answer:

data(golub, package = "multtest")

gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))

shap.data<- apply (golub[, gol.fac=="AML"], 1, function(x) shapiro.test(x)$p.value)

sum(shap.data>.05)

shap.fdr <-p.adjust(p=shap.data, method="fdr")

sum(shap.fdr>0.05) #Pass

sum(shap.fdr<0.05) #Fail

We don't reject null hypothesis for p-values greater than 0.05

Therefore 225 genes fail to 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). Useappropriate 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.

Answer:

H0: The gene "HOXA9 Homeo box A9" expresses at the same level as the “CD33” gene.

HA: The gene “HOXA9 Homeo box A9” doesn’t express at the same level as the “CD33” gene.

data(golub, package = "multtest")

gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))

grep("HOXA9 Homeo box A9",golub.gnames[,2])

grep("CD33",golub.gnames[,2])

wilcox.test (x= golub[1391, gol.fac=="ALL"], y= golub[808, gol.fac=="ALL"],

paired=T, alternative="two.sided")

p-value = 0.01242 which is less than .05. Here the genes express at different levels.Hence, we reject the null hypothesis.

**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 theanswer.

Answer:

H0: gender and admission are independent for each department.

HA: gender and admission are dependent for each department.

library(datasets);

str(UCBAdmissions)

dpt<- c("dpt = A","dpt = B", "dpt = C", "dpt = D", "dpt = E", "dpt = F")

for (i in 1:6 ){

print(dpt[i])

dpt.Data <- matrix(c(UCBAdmissions[1,1,i], UCBAdmissions[2,1,i],

UCBAdmissions[1,2,i],

UCBAdmissions[2,2,i]), nrow=2,

dimnames= list("Admit"=c("Admitted","Rejected"),

"Gender"=c("Male","Female")))

print(dpt.Data)

print(chisq.test(dpt.Data))

print(fisher.test(dpt.Data))

}

Department A:

p = 5.205e-05 is less than 0.05 so, we can reject null hypothesis and hence the Gender and Admissions are dependent for department A. By which we can say that Alternate hypothesis is true.

Departments B, C, D, E, F:

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

All the p-values are greater than 0.05. So, we fail to reject null hypothesis. Hence, the gender and admissions are most probably independent.

**Problem 5: (20 points):**

There are two random samples X1...Xn and Y1...Ym with population means μX and μY and population variances σ^2X and σ^2Y . For testing H0 : σ ^2X = σ^ 2Y versus HA : σ ^2X < σ ^2Y , we can use a permutation test for the statistic S = S^2X/S^2Y.

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.

Answer:

Hypothesis:

H0 : σ ^2X = σ^ 2Y / VAR(ALL) = VAR(AML)

HA : σ ^2X < σ ^2Y / VAR(ALL) < VAR(AML)

**# For CD33:**

install.packages('gtools')

library(gtools)

data(golub, package = "multtest")

gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))

genedata <- grep("CD33",golub.gnames[,2])

golubdata <- golub[genedata,]

n <- length(data)

Test.obs <- var(golubdata[gol.fac=="ALL"]) / var(golubdata[gol.fac=="AML"])

perm = 2000

Test.perm = NULL

for(i in 1:perm) {

data.permtest = sample(golubdata, n, replace=F)

Test.perm[i] = var(data.permtest[gol.fac=="ALL"]) /

var(data.permtest[gol.fac=="AML"])

}

mean(Test.perm<=Test.obs)

Output:

P=0.045. We can reject Null Hypothesis as p is less than 0.05 that var(ALL) = var(AML) for CD33 gene and accept the alternate hypothesis that Var(ALL) < var( AML)

Hence, the variance in the ALL group is smaller than the variance in the AML group.