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.
A)
Rscript:

```
data(golub, package='multtest')
gol.fac <- factor(golub.cl,levels=0:1, labels=c("ALL","AML"))
data<-NULL
for(i in 1:length(golub[,1]))
{
data[i]<-wilcox.test (golub[i,] ~gol.fac, paired=F,
alternative="greater")$p.value
}
genes<-sum(data < .05)
genes
pfdr<-p.adjust(data, method='fdr')
sum(pfdr<.05)
```

Answer:
Genes whose mean mean expression values are higher in ALL group than in AML group is 698

407 genes are expressed higher in ALL group after FDR adjustments

B)
Rscript:

```
data(golub, package='multtest')
gol.fac <- factor(golub.cl,levels=0:1, labels=c("ALL","AML"))
data<-NULL
for(i in 1:length(golub[,1]))
{
  data[i]<-wilcox.test (golub[i,] ~gol.fac, paired=F,
```

```
alternative="greater")$p.value
}
genes<-sum(data < .05)
genes
pfdr<-p.adjust(data, method='fdr')
sum(pfdr<.05)
ordered <-order(data,decreasing="False")
print(golub.gnames[ordered[1:3],2])
orderedfdr <- order(pfdr,decreasing="False")
print(golub.gnames[orderedfdr[1:3],2])

meanALL = apply (golub[, gol.fac=="ALL"], 1, mean)
meanAML = apply (golub[, gol.fac=="AML"], 1, mean)
diff = meanALL - meanAML
orderdiff <- order(diff, decreasing=TRUE)
golub.gnames[orderdiff[1:3],2]
```

Answer:
Top three genes with smallest p values without FDR adjustments:
[1] "TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)"
[2] "Macmarcks"
[3] "VIL2 Villin 2 (ezrin)"

Top three genes with smallest p values with FDR adjustments:
[1] "Macmarcks"
[2] "VIL2 Villin 2 (ezrin)"
[3] "TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)"

Top three genes with largest difference between the means in the ALL group and the AML group
[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"

Top three genes with smallest p values are not same as top three genes with largest difference between the means in 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.

Rscript:

```
data(golub, package='multtest')
gol.fac <- factor(golub.cl,levels=0:1, labels=c("ALL","AML"))
data<-NULL
for(i in 1:length(golub[,1]))
{
  data[i]<-shapiro.test(golub[i,gol.fac=="AML"])$p.value
}
mean(data)
pfdr<-p.adjust(data,method='fdr')
sum(pfdr<0.05)
```

Answer:
P value: 0.3457752
P value is greater than .05 so we fail to reject the null hypothesis, which means gene expression values in AML group follows normal distribution.

225 genes 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.

Rscript:

```
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])
x<-golub[1391,gol.fac=="ALL"]
y<-golub[808,gol.fac=="ALL"]
wilcox.test (x,y, paired=T, alternative="two.sided")
```

Answer:

Wilcoxon signed rank test with continuity correction

data: x and y
V = 62, p-value = 0.01242
alternative hypothesis: true location shift is not equal to 0

Here p value is 0.01242 so we can conclude that we reject the null hypothesis and say that for ALL patients two genes expresses 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.
What are your conclusions? Please submit your R script with the answer.
Rscript:

```
library(datasets)
data(UCBAdmissions)
# 1st Way
dept<-c("Dept A","Dept B", "Dept C","Dept D", "Dept E", "Dept F")
for(i in 1:6){
  print(dept[i])
test.table<-
matrix(c(UCBAdmissions[1,1,i],UCBAdmissions[2,1,i],UCBAdmissions[1,
2,i],UCBAdmissions[2,2,i]), nrow=2,
dimnames=list("Status"=c("Admitted","Rejected"),"Gender"=c("Male","Fe
male")))
print(test.table)
print(fisher.test(test.table))
}
#2nd Way
mat<-apply(UCBAdmissions[ , , ],3, function(x) fisher.test(x) )
mat
```

Answer:

For Dept A, p value is $1.669e^{-5}$ which is less than 0.05 so we reject the null hypothesis and conclude that admission decision and gender is dependent.

For Dept B to Dept F all the p values are above 0.05 i.e. 0.6771, 0.3866, 0.5995, 0.3604, 0.5458 respectively so we fail to reject null hypothesis and concludes that admission decision for Dept B to F is independent of Gender.

```
[1] "Dept A"
          Gender
Status     Male Female
  Admitted  512     89
  Rejected  313     19
```

```
        Fisher's Exact Test for Count Data

data:  test.table
p-value = 1.669e-05
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1970420 0.5920417
sample estimates:
odds ratio
 0.3495628

[1] "Dept B"
          Gender
Status     Male Female
  Admitted  353     17
  Rejected  207      8

        Fisher's Exact Test for Count Data

data:  test.table
p-value = 0.6771
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2944986 2.0040231
sample estimates:
odds ratio
 0.8028124

[1] "Dept C"
          Gender
Status     Male Female
  Admitted  120    202
  Rejected  205    391

        Fisher's Exact Test for Count Data

data:  test.table
p-value = 0.3866
alternative hypothesis: true odds ratio is not equal to 1
```

95 percent confidence interval:
 0.8452173 1.5162918
sample estimates:
odds ratio
    1.1329

[1] "Dept D"
          Gender
Status   Male Female
  Admitted 138   131
  Rejected 279   244

        Fisher's Exact Test for Count Data

data: test.table
p-value = 0.5995
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.6789572 1.2504742
sample estimates:
odds ratio
 0.9213798

[1] "Dept E"
          Gender
Status   Male Female
  Admitted   53   94
  Rejected 138   299

        Fisher's Exact Test for Count Data

data: test.table
p-value = 0.3604
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.8064776 1.8385155
sample estimates:
odds ratio
  1.221185

```
[1] "Dept F"
        Gender
Status    Male Female
  Admitted   22     24
  Rejected  351    317

        Fisher's Exact Test for Count Data

data:  test.table
p-value = 0.5458
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.4332888 1.5756278
sample estimates:
odds ratio
 0.8280944
```

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 $S = \frac{S_x^2}{s_y^2}$.

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.
Rscript:

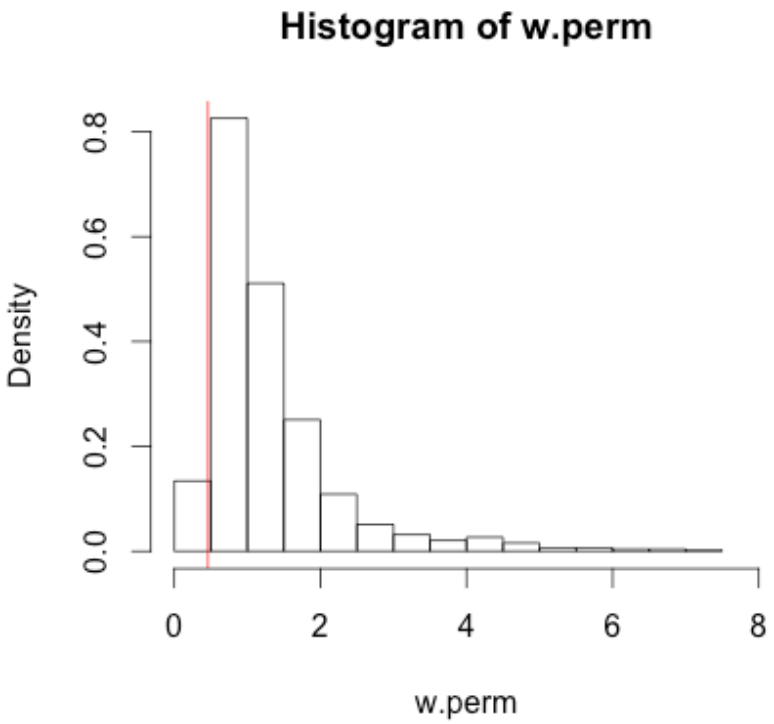
```
data(golub, package = "multtest")
gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))
data<- golub[808,]
n<-sum(gol.fac=="ALL")
m<-sum(gol.fac=="AML")

w.obs<-var(golub[808,gol.fac=="ALL"])/var(golub[808,gol.fac=="AML"])

n.perm=2000
w.perm =rep(NA,n.perm)
for(i in 1:n.perm){
  data.perm = sample(data,n+m,replace=F)
  w.perm[i] = var(data.perm[gol.fac=="ALL"])/var(data.perm[gol.fac=="AML"])
}

mean(w.perm<=w.obs)
hist(w.perm, freq=FALSE, xlim=c(0,8)) #histogram, show proportion instead of
frequence (=FALSE), x range 0 to 30
abline(v=w.obs, col=2)
```

Answer:
Pvalue is 0.044 which is less than 0.05 so we will reject the null hypothesis and conclude that variance in the ALL group is smaller than variance in the AML group for CD33 gene.


Histogram of w.perm
Density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
w.perm